Shifeng Liu · Gábor Bohács · Xianliang Shi · Xiaopu Shang · Anqiang Huang *Editors*

LISS 2020

Proceedings of the 10th International Conference on Logistics, Informatics and Service Sciences

Springer

LISS 2020

Shifeng Liu · Gábor Bohács · Xianliang Shi ·
Xiaopu Shang · Anqiang Huang
Editors

LISS 2020

Proceedings of the 10th International Conference on Logistics, Informatics and Service Sciences

Springer

Editors

Shifeng Liu
Beijing Jiaotong University
Beijing, China

Xianliang Shi
Beijing Jiaotong University
Beijing, China

Anqiang Huang
Beijing Logistics Informatics
Research Base
Beijing Jiaotong University
Beijing, China

Gábor Bohács
Budapest University of Technology
and Economics
Budapest, Hungary

Xiaopu Shang
International Center for Informatics
Research
Beijing Jiaotong University
Beijing, China

ISBN 978-981-33-4358-0 ISBN 978-981-33-4359-7 (eBook)
https://doi.org/10.1007/978-981-33-4359-7

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Mono Camera Based Pallet Detection and Pose Estimation for Automated Guided Vehicles . . . 1
Gabor Bohacs, Zoltan Rozsa, and Balint Bertalan

Modeling Strategies for Risk Prediction in Clinical Medicine with Restricted Data: Application to Cardiovascular Disease . . . 13
Junyoung Lee and Wai Kin (Victor) Chan

Pragmatic Interoperability for Extreme Automation and Healthcare Interoperability and Continuity . . . 31
Jinan Fiaidhi, Sabah Mohammed, and Sami Mohammed

AHP-Based Multicriterial Ranking Model for the City Logistics Analysis of Urban Areas . . . 45
Dávid Lajos Sárdi and Krisztián Bóna

The Influence of Religious Tradition and Social Trust on Corporate Cash Holdings . . . 61
Zhijia Peng, Ming Xiao, and Lin Mu

Data-Driven Organizational Structure Optimization: Variable-Scale Clustering . . . 79
Ai Wang and Xuedong Gao

Auto-carrier Transport Routing Problem with Loading Constraint . . . 91
Xiqing Yang, Wenliang Bian, Hanping Hou, and Xiying Yang

Towards Using Micro-services for Transportation Management Systems . . . 107
Sabah Mohammed, Jinan Fiaidhi, and Mincong Tang

An Analytic Approach for Workers' Fatigue Examination Using RFID-Enabled Production Data . . . 119
Yishu Yang and Ray Y. Zhong

Port Logistics Demand Forecast Based on Grey Neural Network with Improved Particle Swarm Optimization 133
Ruiping Yuan, Hui Wei, and Juntao Li

A Review of the Schedulability Tests and Optimal Design for Rate-Monotonic Scheduling 145
Yang Li, Xiuli Wang, Tianying Liu, Youwei Wang, Jianming Zhu, and Meijiao Duan

The Facility Layout Problem in a Logistics Park Based on Accessibility and Transport Diversity 159
Lu Qin and Yi Zhao

The Study of the Mode and Operation Mechanism of Government Procurement Management in China 175
Bing Hou, Ying Chai, and Xiao Xiao

Non-financial Influencing Factors of Software Enterprise Credit Level 193
Lingling Wang, Shifeng Liu, and Shouting Miao

Evacuation Design and Simulation of Rail Transit Platforms 207
Fang Wang

Optimization of Non-motorized Traffic Around Rail Transit Stations 221
Ziyao Zhao, Li Wang, and Xiaoning Zhu

Layout Planning of Container Piggyback Transport Stations Based on Road-Railway Intermodal Transportation 231
Tian Xia, Xiaoning Zhu, and Li Wang

Simulation and Optimization of Automated Warehouse Based on Flexsim 245
Xiang Li, Li Wang, and Xiaoning Zhu

Research on Path Planning of Logistics Storage Robot Based on Fuzzy Improved Artificial Potential Field Method 263
Guanyi Liu, Yanping Du, Xinyue Li, and Shuihai Dou

The Whole-Course Logistics Enterprise Credit Reference System Based on Blockchain Technology 279
Yang Mao, Minzhen Huang, Shifeng Liu, Yi Song, Guohua Li, and Jingya Liu

Predicting Question Response Rate in Different Topics in Zhihu 291
Hong-de Liu, Ming Li, and Cheng-qi Yang

Content Recommendation of Tender Documents Based on Qualitative Characteristics 305
Tingting Zhou, Guiying Wei, and Ai Wang

Environmental Uncertainty, Customer Concentration and Debt Financing Capacity 323
Jiaqi Ma and Lin Han

OD-HyperNet: A Data-Driven Hyper-Network Model for Origin-Destination Matrices Completion Using Partially Observed Data 335
Yuxuan Xiu, Wanda Li, Jing Yang (Sunny) Xi, and Wai Kin (Victor) Chan

The Supervision Strategy Based on Evolutionary Game Between Government and Sharing Economy Enterprises of Beijing 351
Dandan Li and Xiaochuan Li

Research on Learning Path Recommendation in Intelligent Learning 361
Wenjing Dong and Xuedong Chen

News Recommendation Based on Click-Through Rate Prediction Model 373
Guiying Wei, Yimeng Wei, and Jincheng Lei

Data-Driven Phenetic Modeling of Scripts' Evolution 389
Gábor Hosszú

A Relevance Study of Economic Time Series Data 405
Lei Han, Wei Cui, and Wei Zhang

Deciphering Historical Inscriptions Using Machine Learning Methods 419
Loránd Lehel Tóth, Gábor Hosszú, and Ferenc Kovács

Urban Public Traffic Network Optimization Based on Simulation Technology 437
Juenuo Yang, Li Wang, and Xiaoning Zhu

Special Purpose Network Simulator for Transport Protocol Analysis 447
Dávid Tegze, Ferenc Kovács, and Gábor Hosszú

Financing Schemes in Supply Chains with a Capital-Constrained Supplier: Coopetition and Risk 461
Danqin Yang, Jiani Ding, and Nan Yan

Evaluating Transportation Routes Between China and Vietnam Based on Delphi–CFPR 477
Thi Phuong Thao Doan, Lixin Shen, Xiaowen Shi, Zaili Yang, Chuang Li, and Ke Jing

Integrated Planning for the Intermodal Container Terminals of the CR Express based on Markov Process 495
Cuijie Diao, Shujuan Guo, Gang Li, Xiaohan Wang, and Zhihong Jin

Research on Quantitative Stock Selection Method Based on Random Forest 509
Haining Yang and Xuedong Gao

MPNC: Improving Reliability of Data Transmission in Wireless Ad Hoc Networks 523
Mingming Yan, Xu Li, and Wenjun Huang

Multi-step Time Series Forecasting of Bus Passenger Flow with Deep Learning Methods 539
Feng Jiao, Lei Huang, and Zetian Gao

Research on Multi-stage Random Distribution Method 555
Fachao Li, Yueye Zhang, and Chenxia Jin

Vessel Scheduling and Refueling in Liner Shipping: Modeling Transport of Perishable Assets with a Collaborative Agreement 567
De-chang Li, Hua-long Yang, and Wei Song

Research on the Common Delivery Model of Express Logistics in Urban and Rural Areas—A Case Study on Lujiang County, Hefei City 583
Zhou Yao, Shuihai Dou, Guanyi Liu, and Yanping Du

Analysis of Lemon Company's Cross-Border E-Commerce Logistics Distribution Mode Selection 601
Li Qin Hu, Amit Yadav, Hong Liu, Sami Azam, Asif Karim, Bharanidharan Shanmugam, Abdul Hasib Siddique, and Mehedi Hasan

The Construction of "Two-Side" Value Chain Between the Power Grid Side and the User Side in the Internet Trading Platform with Positive Sum Game Nature 617
Xin Yin and Shaozhen Hong

Mining and Analysis of Emergency Information on Social Media 627
Dan Chang, Lizhu Cui, and Yiming Sun

Application of Object—Attribute Space Segmentation in Bidding Activities 649
Yijie Yin, Yuwen Huo, and Yaoyu Hu

Research on Bidding Case Recommendation Algorithm Considering Bidding Features 665
Wenting Liang, Hao Liu, and Yaoyu Hu

Pricing and Versioning Strategy for Information Products with No-Free-Disposal: The Role of Consumers' Expectations Formation 677
Danqin Yang and Juan Bai

The Comparison of Two Styles of Residence Layout Examined by Sunshine—Beijing and Barcelona 695
Yunan Zhang, Chun Ji, Shuxuan Xing, and Hui Zhou

Manpower Planning for MTR Carriage Assembly with an Integer Programming 703
Yingjiang Wu and Ray Y. Zhong

Relationships Between Human Capital, Organizational Integration, and Performance in Omni-Channel Retailing 717
Yixuan Han, Hanchen Li, Haochen Sun, and Guang Song

Research on the Cashback Strategy of E-tailers with Strategic Consumers 731
Danqin Yang, Nan Yan, and Jiani Ding

Correlation Analysis of Waterway Passenger Transport and Tourism Development in Beihai City 747
Xiaolei Liu, Xiaofeng Li, Wenpeng Fei, Qilong Huang, and Yin Li

Towards Fully-Synthetic Training for Industrial Applications 765
Christopher Mayershofer, Tao Ge, and Johannes Fottner

Research on the Prediction of Shenzhen Growth Enterprise Market Price Index Based on EMD-ARIMA Model 783
Tianhua Li, Shaowei Qu, and Gaoping Huang

Research on Location and Path Planning of Distribution Center Based on Improved k-Means Clustering Algorithm and Improved Ant Colony Algorithm 797
Shuihai Dou, Zhou Yao, Xiaotong Shi, and Guanyi Liu

Customer Churn Prediction in the Broadband Service on Machine Learning 811
Yujie Guo and Lei Huang

Research on the Identification and Prediction of Default Risk of Online Lending Platform Customers 823
ShuaiQi Liu and Sen Wu

Stackelberg Game Model of Power Plants and Large Users Considering Carbon Trading 837
Xinyu Niu and Shifeng Liu

Collaborative Governance of Internal Control of Scientific Research Funds in Colleges and Universities Under the Background of "Release Management Service" 849
Tian Qi, Xue-wei Li, and Jing Li

Multiobjective Optimization of Production Line Supply Based on Maximum Endurance Time and Rest Allowance 863
Sezen Korkulu and Krisztian Bóna

Study on Site Selection Evaluation of Police Drone for the Disposal of Abnormal Moving Vehicles 875
Zhaowei Ding, Xin Wang, Lei Wang, Jianhua Yang, and Ai Wang

Selection Strategies on Overseas Logistics Nodes for Chinese Business Logistics Enterprises 893
Minjie Liang, Zikui Lin, and Kai Zheng

Evaluation Model Construction of Brand Media Influence in Short Video in the Publishing Media Industry 913
Shaozhen Hong and Liang Wang

Balancing Optimization of Mixed-Flow Assembly Line Based on Hybrid Genetic Algorithm 931
Meng Li and Dan Chang

Informatization Development Mode of Internal Control in Colleges and Universities Based on Principal-Agent 947
Yuzhen Cheng

Research on Balancing Problem of Stochastic Two-Sided Mixed-Model Assembly Lines 963
Beibei Zhang and Dan Chang

Intralogistics Conversion of the Physical Internet 981
Eszter Puskás and Gábor Bohács

A Data-Driven Framework for Exploring the Spatial Distribution of Industries 995
Huifeng Sun

The New Data-Driven Newsvendor Problem with Service Level Constraint 1009
Yuqi Ye and Xufeng Yang

Order Acceptance and Scheduling for Make-to-Order Manufacturing Enterprises 1025
Mingzhen Yu

Bayesian Production/Inventory Competition with Unobserved Lost Sales 1037
Shuang He, Pujie Shi, and Jian Zhang

Mono Camera Based Pallet Detection and Pose Estimation for Automated Guided Vehicles

Gabor Bohacs, Zoltan Rozsa, and Balint Bertalan

Abstract Detection and pose estimation of pallets are critical phases in the operation of automated guided vehicles. In this paper, we introduce a novel pipeline for accurate localization based on only a single camera. We utilize the popular YOLO detector, object, and camera models to achieve better performance than the state of the art techniques.

Keywords Detection · Pose estimation · AGV · Automation · Material handling

1 Introduction

Moving pallets (with or without a load on it) is a common task of Automated Guided Vehicles (AGVs) in industrial and logistic systems. To do the transportation, starting with the lifting of the pallet, precise docking to it is inevitable, which requires knowing its exact location. In many cases, pallets are stored in a shelf system or have a predefined location, so the ego position of the AGV could be enough information to navigate the vehicle to the correct position. One reason for the application of the pallet detection and pose estimation pipeline is that the main navigation of the vehicle can be less accurate if the docking is supported. More importantly, there can be numerous cases when the pallet should be transported from an unspecified position, or the position and/or orientation is just approximately known (e.g., the last worker was not careful enough during the deposition). Thus, the flexibility of

G. Bohacs (✉) · Z. Rozsa · B. Bertalan
Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics Budapest, Budapest, Hungary
e-mail: gabor.bohacs@logisztika.bme.hu

Z. Rozsa
e-mail: zoltan.rozsa@logisztika.bme.hu

B. Bertalan
e-mail: blintber@gmail.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_1

a production line with automated material handling cannot be sustained without machines provided with the intelligence of pallet localization. In this paper, we propose a novel solution to this problem that offers better performance than state of the art and requires only a single camera.

1.1 Contributions

The paper contributes to the following:

- A new methodology is proposed to localize pallets (with load) in an undefined position in an industrial environment.
- For the operation, only a single camera is required, but the 6D pose is estimated.
- Comparison of a different detection algorithm for pallet detection purposes is provided.
- Measurements for evaluating the proposed pipeline in different scenarios in terms of angle and distance are provided.

1.2 Outline of the Paper

The paper is organized as follows: Section 2 surveys the literature about the related topics. Section 3 compares the state of the art object detection methods for pallet detection purpose, while Section 4 describes the proposed pose estimation method and the whole pipeline in detail. Sections 5 shows our test results and comparison to other methods. Finally, Section 6 draws some conclusions.

2 Related Works

The scientific community of robotics and automated material handling is interested in pallet localization for decades. Garibott et al. [1] used a camera for detection and sonar sensor for pose estimation.

Various depth sensors are used for determining the position and orientation of pallets, like stereo camera [2, 3], Time of Flight (ToF) cameras [4, 5] or Kinect [6, 7], but the main sensors used in this topic are range scanners [8, 9]. The advantage of 2D LIDARs compared to other sensors is the less information contained in one frame (less complex search). However, their main disadvantage also lies in this property, as only a plane can be examined, the pallets cannot be located at different heights. State of the art solutions with these devices utilizes the latest results of machine vision like Convolutional Neural Networks (CNN). These techniques offer high performance but require a large amount of training data [10]. For example [11] applied Faster R-CNN [12] for detection. Nevertheless, relying on depth sensors is a drawback from a given aspect, as they are generally more expensive than a camera, which is also already part of the sensor system to give support in the detection phase.

Several methods utilize only a mono camera for the detection and pose estimation based on some geometric modeling. Pages et al. [13] and [14] used pinhole camera model, Tsai [15] calibration, and the fact that the pallets lie on the ground plane, [14] also simplified the extrinsic estimation by using a camera with parallel z axis to the ground floor. Byun and Kim [16] utilizes a virtual plane and back projection of pallet corners to it to find first the angle of the pallet and then the distance to it in their fork coordinate system. Cui et al. [17] also uses Tsai coplanar calibration, pallet model, but only two points (two corners from the edge between the ground and the pallet) are utilized to estimate the orientation and center point of the pallet (we will use too). These methods utilize the same geometric properties as we. Still, our model has the advantage of having one more degree of freedom, as we can determine the offset of the pallet plane in z direction (the other ones exclude the possibility of localization above grounds, e.g., on shelves).

3 Detection

Pallet detection solution can utilize fiducial markers [18] to facilitate detection. Still, they can become contaminated or damaged, and additional effort is required to equip all the actual and future pallet stock.

Thus detection usually is based on basic image processing like color space transformation to reduce the effect of different lightings, thresholding [14], and morphology operations [8] for BLOB (Binary Large Object) detection or edge detection and Hough transformation [17] for finding the contour of the pallet. These operations have the advantage of low complexity, but they are sensitive to noise, so their performance is below the methods designed especially for detection.

Conventional detection algorithms are used in pallet recognition too, they are based on so-called handcrafted features, like Haar-like features [19], ICF (Integral Channel Features) [2], ACF (Aggregate Channel Features) [3] or features designed specially for pallet detection [20]. These types of methods are slower but have better performance.

CNN based detectors [21] are the current state of the art of image-based object detection. So, we decided to test these for pallet detection tasks. The expectation is that they can be slower than conventional detectors, but can perform even better. We compared BLOB detection based pallet identification, ACF detector with two states of the deep learning based detector. These two are Faster R-CNN [12] and YOLO (You Only Look Once) [22] networks. The result of the tests can be seen in Table 1. The evaluation is based on our test database, which contains 120 images; each image with one pallet, sample images are shown in Fig. 1. The training database was constructed from 180 images, a relatively small number for neural network-based

Table 1 Comparison of detection methods for pallet detection purpose

Method	BLOB	ACF	Faster R-CNN	YOLO
Training time required (s)	0	799	18,375	26,123
Running time (ms)	26.94	120.5	1224.31	267.54
Recall	0.45	0.6	0.99	0.98
Precision	0.95	0.7	0.72	0.99
F-measure	0.61	0.65	0.83	0.99

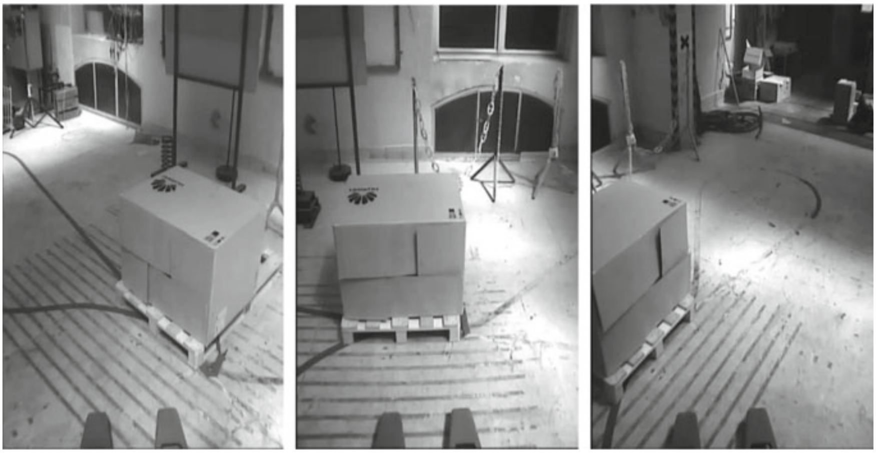
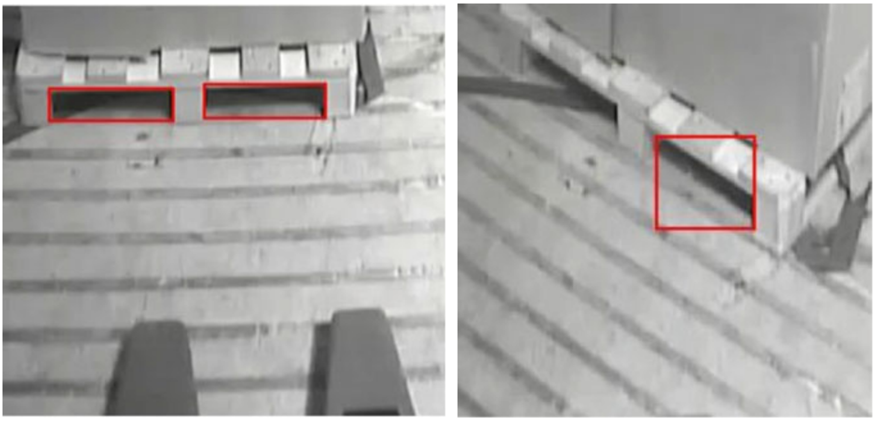


Fig. 1 Example images of our dataset

training. That is why we used data augmentation to multiply the number of images (the original images also had extreme pallet positions and orientations). Our detector was trained to find the two holes of the pallet; in this way, it is more robust against only partially visible objects. Examples of detection results can be seen in Fig. 2.

Our tests were run on a laptop computer with the following configuration: Intel Core i7-6700HQ CPU, 8 GB RAM, NVIDIA GeForce GTX 960 M GPU, and Windows 10 operating system. Our test results show that Faster R-CNN and YOLO are the most effective methods for the pallet detection problem, as the YOLO is much faster (also better in overall performance), we applied this in our pipeline. In [20] and [3], the authors showed that their features designed for pallet detection could achieve better performance on their database than ACF. Still, on the one hand, the reported running time is too high (more than 1 s in similar image size). On the other hand, they uses stereo camera for pallet candidate generation, so we did not consider this method for our real-time operation, which also requires an additional sensor.



(a) Successful BLOB detection

(b) Unsuccessful BLOB detection

Fig. 2 Detection on example images

4 Pose Estimation

In this section, we introduce the proposed pose estimation from a single camera image. Our assumptions are the following:

- The camera has a fix position at the forklift.
- We have a fully calibrated camera (intrinsic and extrinsic) [23].
- Transformation from the camera coordinate system to the forklift coordinate system is known.
- The equation of the ground plane is known in the forklift coordinate system (in our case, the origin of this coordinate system is located on the ground plane and its normal vector oriented toward direction z).
- The pallet we found (in the previous detection step) lies either on the ground or a plane parallel to it (shelf system).

If the assumption above holds, only two key points in the image from our pallet and the real distance between them (known from the models of standard pallets) are enough to determine the orientation of the pallet and the 3D position of its center point. Naturally, if more than two correspondences can be established, the solution can be refined. The solution is illustrated in Fig. 3. The offset of the plane (in the direction of its normal vector) containing the points can be determined if the distance between the points is known. Giving solution based only on the minimum, two points are important because, in the case of multiple pallets. Then robust model fitting (like RANSAC—RANdom SAmple Consensus [24]) can be used, and the complexity of these algorithms depends on model parameters.

The equations in the coordinate system of the forklift:

$$sm'_1 = A[R|t]M'_1 \tag{1}$$

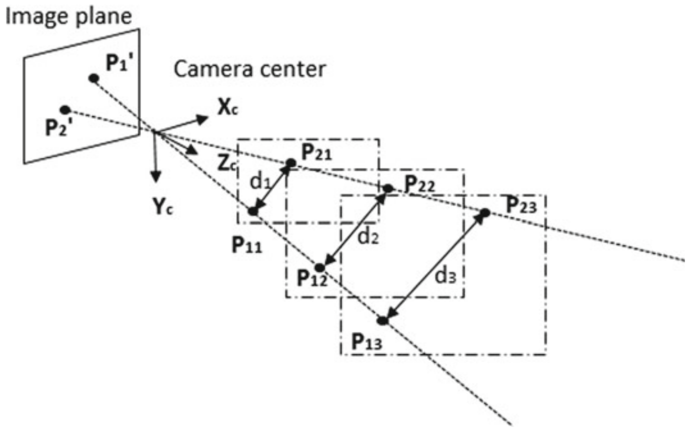


Fig. 3 Illustration of the solution. P_i' are the image of the points and P_{ij} are the 3D coordinates if the distance is between them is d_j

$$sm_2' = A[R|t]M_2' \tag{2}$$

$$\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} = d \tag{3}$$

$$Z_1 = Z_2 \tag{4}$$

where s is an arbitrary scale factor $m_i = [u_i v_i\ 1]^T$ the coordinates of the projection on the image plane of $M_i = [X_i Y_i Z_i\ 1]^T$, A is the matrix of intrinsic parameters, $[R|t]$ describes the camera extrinsic [23]. Equation 1 and 2 are matrix equations of the projection of two points in pinhole camera model, which can be rewritten in a form of 4 scalar equations. The unknown variables are the elements of M_i, real world coordinates of the two points.

After solving the six scalar equations for the six unknown, the angle difference between the pose of the forklift and the pallet can be calculated, as the direction vector of pallet front plane can be defined as:

$$v = \frac{X_2 - X_1}{d} i + \frac{Y_2 - Y_1}{d} j \tag{5}$$

Illustration of different coordinate systems can be seen in Fig. 4.

The coordinates of the pallet front center:

$$X = \frac{X_1 + X_2}{2}, Y = \frac{Y_1 + Y_2}{2}, Z = Z_1 = Z_2. \tag{6}$$

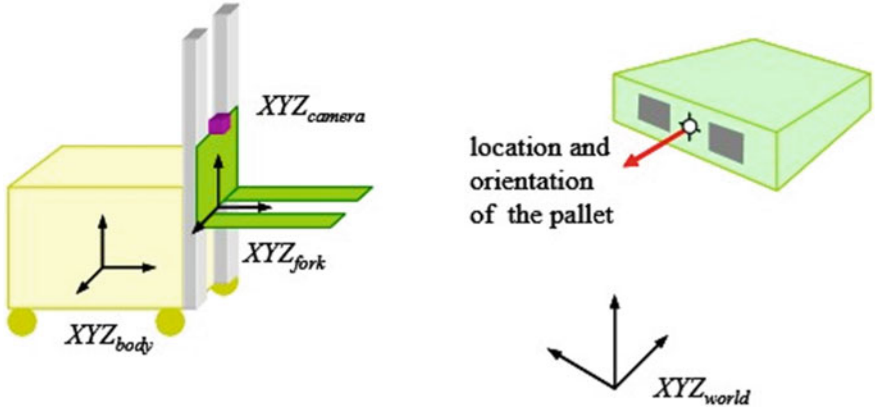


Fig. 4 Illustration of different coordinate system definitions in [16]. We used similar coordinate system definitions. In our case the forklift coordinate system is fixed to the ground (moving world coordinate system) and the camera is in a fix position on the AGV (it is not moving with the fork)

5 Test Results

Our tests have been carried out on the automatized forklift visible in Fig. 5. A SICK NAV350 sensor provides the navigation of the vehicle and the ground truth data.

We did several evaluations to measure the accuracy of our pose estimation based on a single camera and YOLO detection. We tested the proposed pipeline in different lighting conditions (natural and artificial). The pose estimation tests are based on 16 ground truth measurement, where the pallet distance varied from about 2000–2300 mm and the angle (enclosed by the normal vector of the pallet front and the direction vector of the AGV) from −60° to 60°, The average distance error is **48.12 mm**, and the average angle error is **1.48**°. The results are illustrated in Figs. 6 and 7.

Relatively high error values occur when a large angle is enclosed by the vehicle and the pallet, which is unlikely in real-life situations. Our pose estimation test results show very high accuracy for a single camera. It is comparable to the results achieved in other works even with stereo camera [2].

6 Conclusion

In the paper, a pipeline is proposed to localize pallets utilizing only a single camera. The methodology is novel in many aspects.

The proposed method finds the pallet if only its front plane (even is some case just partially) visible. At the same time, many of the earlier attempts can only detect the pallet if much more information is available (so excluding the pallet localization

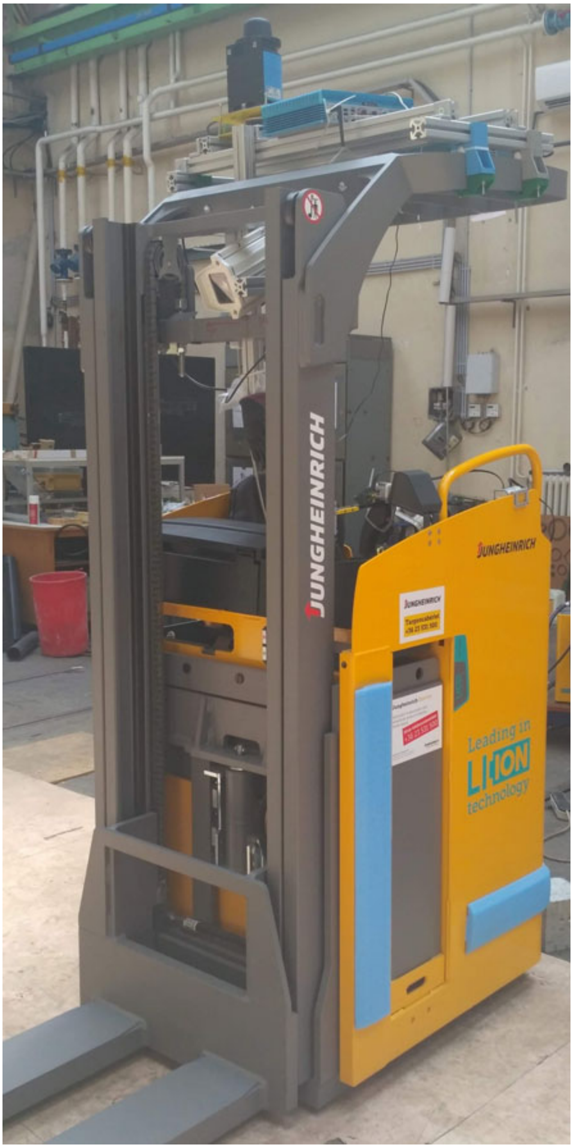


Fig. 5 Automatized forklift with fixed position camera

possibility in case of load on it). We estimate the 3D position of the object compared to other mono camera based pose estimation, where only 2D position is estimated to the best of our knowledge. Our tests prove that the pipeline outperforms the state of the art solutions for this problem. In the future, we plan to further speed up the process by extending the pipeline with tracking.

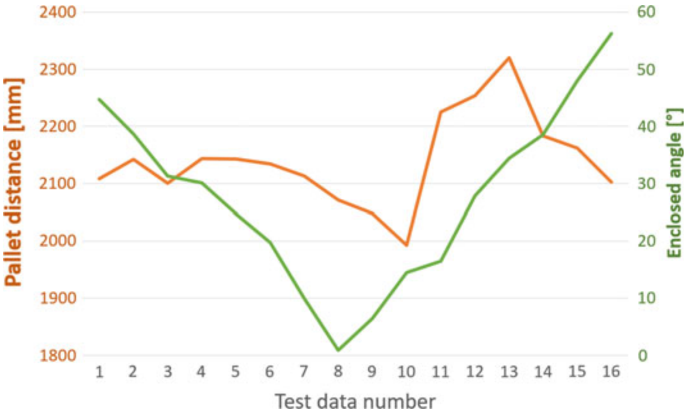


Fig. 6 Variation of 2D pose in case of pose estimation tests

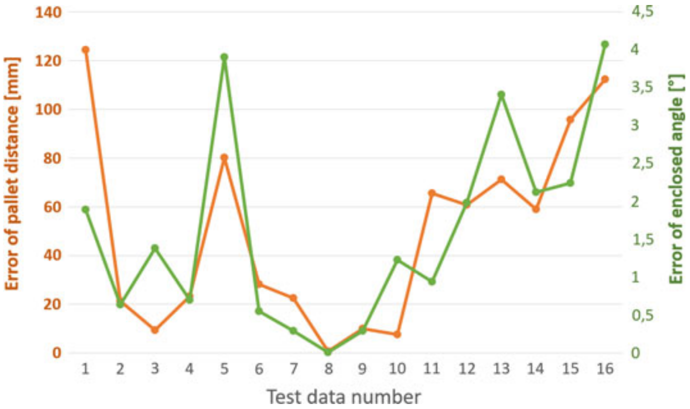


Fig. 7 Accuracy of pose estimation tests

Acknowledgements Project no. 2017-1.3.1-VKE-2017-00036 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the VKE_17 funding scheme.

References

1. G. Garibott, S. Masciangelo, M. Ilic, P. Bassino, Robolift: a vision guided autonomous fork-lift for pallet handling, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS '96*, vol. 2, pp. 656–663 (1996)
2. R. Varga, S. Nedevschi, Vision-based autonomous load handling for automated guided vehicles, in *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 239–244 (2014)

3. R. Varga, A. Costea, S. Nedevschi, Improved autonomous load handling with stereo cameras, in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 251– 256 (2015)
4. D. Haanpaa, G. Beach, C.J. Cohen, Machine vision algorithms for robust pallet engagement and stacking, in *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–8 (2016)
5. K. Terabayashi, I. Takashima, Y. Suzuki, A. Hinami, T. Sasaki, Easy acquisition of range image dataset for object detection using retroreflective markers and a time-of-flight camera: An application to detection of forklift pallets, in *Proceedings of the Seventh Asia International Symposium on Mechatronics* (B. Duan, K. Umeda, and W. Hwang, eds.), (Singapore), pp. 1001–1005, Springer Singapore (2020)
6. J.-Y. Oh, H.-S. Choi, S.-H. Jung, H.-S. Kim, H.-Y. Shin, An experimental study of pallet recognition system using kinect camera, pp. 167–170 (2013)
7. J. Xiao, H. Lu, L. Zhang, J. Zhang, Pallet recognition and localization using an rgb-d camera. Int. J. Adv. Rob. Syst. **14**, 172988141773779 (2017)
8. N. Bellomo, E. Marcuzzi, L. Baglivo, M. Pertile, E. Bertolazzi, M.D. Cecco, Pallet pose estimation with lidar and vision for autonomous forklifts, in *IFAC Proceedings Volumes*, *13th IFAC Symposium on Information Control Problems in Manufacturing*, vol. 42, no. 4, pp. 612–617 (2009)
9. L. Baglivo, N. Biasi, F. Biral, N. Bellomo, E. Bertolazzi, D. Lio, M. Cecco, Autonomous pallet localization and picking for industrial forklifts: A robust range and look method. Measur Sci Technol **22**, 085502 (2011)
10. S. Mohamed, A. Capitanelli, F. Mastrogiovanni, S. Rovetta, R. Zaccaria, A 2D laser rangefinder scans dataset of standard eur pallets. Data Brief. **24**, 103837 (2019)
11. S. Mohamed, A. Capitanelli, F. Mastrogiovanni, S. Rovetta, R. Zaccaria, Detection, localisation and tracking of pallets using machine learning techniques and 2D range data. Neural Comput. Appl. (2019)
12. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards realtime object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 06 (2015)
13. J. Pages, X. Armangué, J. Salvi, J. Freixenet, J. Marti, A computer vision system for autonomous forklift vehicles in industrial environments **1**
14. G. Chen, R. Peng, Z. Wang, W. Zhao, Pallet recognition and localization method for vision guided forklift, 1–4 (2012)
15. R. Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J. Rob Autom **3**, 323–344 (1987)
16. S. Byun, M. Kim, Real-time positioning and orienting of pallets based on monocular vision, in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, vol. 2, pp. 505–508 (2008)
17. Cui, G., Lu, L., He, Z., Yao, L., Yang, C., Huang, B., Hu, Z. A robust autonomous mobile forklift pallet recognition, in *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*, vol. 3, pp. 286–290 (2010)
18. M. Seelinger, J. Yoder, Automatic pallet engagement by a vision guided forklift, in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 4068–407 (2005)
19. J.-L. Syu, H.-T. Li, J.-S. Chiang, C.-H. Hsia, P.-H. Wu, C.-F. Hsieh, S.-A. Li, A computer vision assisted system for autonomous forklift vehicles in real factory environment. Multimedia Tools Appl **76**, 11 (2016)
20. R. Varga, S. Nedevschi, Robust pallet detection for automated logistics operations, in *VISIGRAPP* (2016)
21. T. Li, B. Huang, C. Li, M. Huang, Application of convolution neural network object detection algorithm in logistics warehouse. J. Eng **2019**(23), 9053–9058 (2019)

22. J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525 (2017)
23. Z. Zhang, A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000)
24. P. Torr, A. Zisserman, Mlesac: A new robust estimator with application to estimating image geometry. Comput. Vis. Image Underst. **78**(1), 138–156 (2000)

Modeling Strategies for Risk Prediction in Clinical Medicine with Restricted Data: Application to Cardiovascular Disease

Junyoung Lee and Wai Kin (Victor) Chan

Abstract This paper describes modeling strategies for risk prediction in clinical medicine, mainly with respect to survival analysis. Restricted data, which is commonly given in initial clinical research, is assumed for these strategies. Cox's proportional hazard model is used with modern statistical approaches. In this paper, detailed modeling strategies for clinical risk prediction are proposed and demonstrated by using a case study on the cardiovascular disease. Experiments were conducted by employing Stepwise selection and Elastic Net with bootstrapping. Results give some insights for risk prediction and modeling with limitation of clinical data.

Keywords Clinical risk prediction · Cox's proportional hazard model · Bootstrapping · Variable selection · Stepwise selection · Elastic net

1 Introduction

Recent era of information technology and artificial intelligence has increasingly required information system in medical services. More and more hospitals and medicine companies would like to establish and make full use of such a system [1]. The utilization of data, nonetheless, is limited by applications. One of main reasons is demanding reliability of such applications with medical expertise, which could discourage both medical practitioners and data analyst from using the information system. This paper focuses on a guideline of developing prognostic models in clinics, for people in both fields. Especially, this paper introduces modeling strategies of clinical risk prediction. While clinical risk prediction has been widely studied, what differentiates the present work from existing studies is that this paper assumes scarcity of data, which is a common problem appeared in most clinical dataset due

J. Lee · W. K. Chan (✉)
Tsinghua-Berkerley Shenzhen Institute, Tsinghua University, Shenzhen, China
e-mail: chanw@sz.tsinghua.edu.cn

J. Lee
e-mail: lijuanro18@mails.tsinghua.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_2

to rare events or difficulty in data collection. This scarcity problem of course has been identified in field [2]. Building on existing studies, the present paper introduces a holistic approach to deal with this problem and illustrates its application on a case study.

If risks, waiting for us anywhere and anytime in any different domains, were manageable, it would be no longer threating us. Clinical risk management is momentous because sudden diseases can destroy our lives. There has been several researches on estimating the clinical risk of specific death or sudden diseases, such as heart attack or stroke. It is practical in clinics to discriminate high-risk patients and give detailed instruction based on the risk factors. As well, it has an educational function to patients, showing a pressing danger in front of their eyes. However, most studies have been done on restricted circumstances (such as cohorts in specific countries or different era). For this reason, it is encouraged to have clinical risk prediction model in each medical center, referred to previous researches and designed on its own purposes. This allows generalization of such models and can contribute to the research domains by providing prior knowledge in different conditions. Risk in clinic is usually defined in time period after getting in. Thus it is necessary to consider it as *time to event*. Survival analysis is the statistical methods to cover this kind of data. For integrating risk factors in the model, Cox's proportional hazard model [3] is the baseline regression model we have used.

Regression modeling is fundamental in many different research backgrounds. and is preferred because of simplicity and reliability. However, from data collection to evaluation, it is a great deal of work to make insightful regression model serving its purpose. When it comes to clinical trials, study design for prediction model could largely affect its performance and efficiency. In particular, it becomes more difficult to model when the number of candidate variables is excessive in given data size. In medical field, it frequently occurs such as genomics. To select important variables and estimate the effect of them accurately, we need to combine statistical approaches and domain knowledge. There are plenty of literatures about clinical regression modeling and they were summarized in this paper with our experience in modeling clinical risk prediction for cardiovascular sample. Core problem is dealing with overfitting and keeping clinical values. The model should be optimized in the limited statistical power of dataset. Careful attention should be paid through modeling steps, which is introduced later on. Especially, typical variable selection techniques, stepwise selection and Elastic Net, will be compared and analyzed in case study. And resampling method played a pivotal role in estimating distribution of the results. These instructions and results were focused on survival data and would be helpful to get some insights for researchers who intend to start risk prediction modeling with survival setting.

This paper is divided into four parts including this introduction section. Section 2 explains methodology of Survival Analysis. And modeling strategies for clinical risk prediction are listed on top of that. Section 3 uses a case study on cardiovascular disease. Finally, Section 4 summarizes the insights and limitations.

2 Methodology

2.1 Clinical Risk Prediction by Survival Analysis

In clinical field, interested events (e.g. death, occurrence and development of certain disease) have been usually recorded and analyzed in *Time to Event (so-called "Survival") data.* Its characteristic is distinct from continuous or binary data we often encounter. Firstly, *censoring* is common because observation time required to notice the event of each subject is restricted for a variety of reasons (e.g. leave of patients, budget and time limits of study design). This censoring makes it hard to apply standard regression for failure time. Also, it should contain two columns to consider, one of which is binary indicator c_i (censored or event) and the other is time variable t_i (until censored or event) in continuous way. For instance, patients could not be simply treated in binary indicator without event occurring time after entrance. It may cause information loss, which is valuable in many clinical situations. Therefore, we could clarify unique features of survival outcomes d_i as (1).

$$d_i = (c_i, t_i) \tag{1}$$

Survival Analysis, one branch of Statistics, is usually employed to deal with this kind of problems not limited to clinical field, but in Engineering, Social Science and Economics. For *Time to Event data*, it defines representative functions such as survival function and hazard function. Survival function (2) literally denotes probability of survival (event-free) at least until the time points. And hazard function (3) is conditional probability of event in very short period around the time point unless the event has occurred before then. By using these functions, we can effectively represent the *Time to Event data.* Like other statistical analysis, there are diverse approaches for Survival Analysis such as hypothesis testing, estimation and so on. In particular, multivariate models for association between *Time to Event data* and factors have been developed by previous scholars. Recently rising trends of Machine Learning techniques such as Support Vector Machines, Recursive Partitioning and Neural Network have contributed to the development of innovative models. However, as shown in [4], the cutting-edge Machine Learning models even with heterogeneous ensemble technique have not achieved satisfactory performance compared to traditional models at the cost of interpretability. It is the reason why most of clinicians and biostatisticians still prefer traditional statistical models.

$$S(t) = \Pr(T > t) \tag{2}$$

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \rangle t)}{dt} = -\frac{S'(t)}{S(t)} \tag{3}$$

"Hazard ratio" is the popular concept to describe survival data, which can be derived from ratio between hazard functions. For multivariate model, this ratio has been assumed to be proportional between different sets of covariates, so called "proportional hazard model". Cox proportional hazard model (Cox model) [3] is one of the most widely used in clinical researches. Based on proportionality assumption of hazard functions, it shows log-linear relationships between hazard ratios and multiple features of patients as (4), where x_i and b_i is ith variable and coefficient each. And $h(t|x)$ represents hazard function conditional to features of each subject.

$$\log \frac{h(t|x)}{h_0(t)} = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{4}$$

Cox model has been established and developed on the basis of robust statistical background. Similar to other regression models, the parameters of the model could be estimated by maximum likelihood method needless to define baseline hazard function $h_0(t)$, the hazard function with setting all variables zero. Its interpretation and analysis is easily made as any other multivariate models. Numerous hypothetical tests can be simultaneously conducted by the model in the situation with confounders. As well, residual tests and graphical analysis could check assumptions on the model which is quite crucial in statistical modeling. Furthermore, the model can be extended and transformed for better applications in different purposes, including nonlinearity, time-dependency or competing risks. Therefore, it has been widely adopted by clinical practices and comprehensive medical fields due to its simplicity, reliability and scalability.

Especially for clinical risk prediction, it is prevalent to use Cox model. In the case of cardiovascular diseases, clinicians have tried to predict the risk of each patient who came in clinics to prevent sudden occurrence of diseases (stroke, myocardial infarction, or other cardiovascular disease). Framingham Study is the milestone study for stroke risk prediction. Since then, several following researchers have designed different cohort studies in different areas, times and targets for clinical risk prediction of cardiovascular attack. Previous studies are summarized in Table 1.

All above studies have made contributions on figuring out risk factors and making reliable prediction to the targeted subjects. Nevertheless, until recently, most of studies depended on clinical insights to build the models rather than making use of standardized and modern statistical methodology. While trying to establish a model

Table 1 Previous studies of risk prediction in cardiovascular disease

Names	Year	Journal	Target
Framingham study [5, 6]	1991/2017	Stroke/circulation	Stroke
SMART study [7]	2013	Heart	Recurrent vascular event
Qstroke study [8]	2013	BMJ	Stroke
China PAR study [9]	2019	Stroke	Stroke

for clinical risk prediction of a cardiovascular disease, we will introduce generalization of modeling process with cutting-edge approaches taking account of clinical aspects further.

2.2 Modeling Strategies for Clinical Risk Prediction

Clinical predictive model specially places importance on reliability of the model because it is directly related to one's health and life. It would be the most preferable to design extensive sampling method of patients for the specific target of prognosis. But it may need tremendous time and cost to execute the study, so it is necessary to define the sample as clearly as possible at the starting point. For clinical prognostic model, it is common to use regression model to figure out the impact of risk factors. However, regression model is never able to achieve perfect model fit for every situation, so we need to optimize the model with available dataset. Following steps would describe the generalized ways to optimize the model grounded on suggestions of previous literatures [10–12] and trials and errors to apply recent techniques. Even though it could be a helpful reference, model builders have to think their situations thoroughly for detailed strategies. All the processes are about making balance between parsimonious/robust model and complex/high-performed model.

1. *Clinical consideration before modeling*

Within the dataset by which one is interested in establishing clinical risk prediction, we should first specify targeted subjects and risky event, as well as potential risk factors among the features of patients. The subjects in the model should be clinically meaningful to have possibility to get disease or death with certain criteria, rather than general population. And it would be recommendable to confine a specific risky event. For example, in SMART study [7], they established the model aimed at risk prediction of recurrent vascular event among patients with arterial disease. Here, concern is scarcity of events in the data or difficulty of discrimination. Some of previous researches combined several events for a target variable or did not distinguish primary or recurrent events in the model. But it may cause misspecification of model by neglecting model assumptions and confusing the target variable. Finally, candidate list of risk factors ought to be set beforehand. Literatures in related field is the best reference for selecting the factors. The list must contain at least similar factors in researches until recently. And additionally feasible factors in the dataset could be included on top of that. Nonlinearity or interactions among the factors also should be taken into account with clinical viewpoint. Interactions between risk factors with age or treatment have been often regarded as important factors. And some variables such as BMI (Body Mass Index) give harmful impact at both high and low values. Usually, cubic polynomials or Restricted Cubic Splines (RCS) have been employed to describe nonlinear association with continuous variables. The most crucial point here with clinical view is to yield appropriate sample size (especially, number of events in binary or survival data) and effective degree of freedom (edf) for the model. If we don't clear up this point, it would be much hard time during further process. There have been suggested about effective sample with formula in each model. By rule of thumbs, it is proper to cover 10–20 effective samples per 1 edf including nonlinear terms and interactions. But many clinical researches still suffer lack of targeted events compared to features of patients, so further steps and case study may guide some points to reduce dimension of risk factors and maximize the utility of given data for the model.

2. *Preliminary steps with raw data*

Before the begin of modeling, we should deal with the clinically potential risk factors and related risk event. It is a quite important step because it is directly associated with model performance and applicability. Clinical consideration above should be reflected in this step, so as to get an adaptable model from the dataset.

Firstly, scarcity of events in categorical variables, missing values and outliers should be investigated, otherwise, it can affect greatly to the model. Some categories in categorical variables may include too small number of patients who experienced the event or even not include any event. In this case, the coefficients of model will be abnormal values or infinity. So we can combine adjacent category with the category with rare events by careful consideration. As well, clinical data inherently contains a lot of missing values, thus they should be properly treated with missing patterns (usually assumed as missing at random). Dropping the missing data could be one way, but cause biasness and losing efficiency. We can truncate features if they have a large amount of missing values, which bring tiny information about the features. Also, just putting in specific value for missing values, such as inserting no family history of disease to make the feature conservative, might be another way, but needs prudent decision on the clinical aspects. In most situation, imputation do work pretty well to cope with missing. Many mechanisms for imputing values, from simply putting median or mode of each feature to complex algorithms such as maximum likelihood method or tree-based method, have been introduced to keep the efficiency of data and reduce biasness by missing values. Recently, multiple imputation is getting the limelight in clinical data analysis, which imputes several times with up-to-date algorithms. Sterne et al. [13] is brief guide about multiple imputation. Extreme values frequently appear in medical data, due to measurement, recording errors or other reasons. Outliers are often defined as values above three times of interquartile range (the range between 1st and 3rd quartiles) above (or below) the 3rd (or 1st) quartile. Like missing values, removal of the values is usually not right way. It is better to treat outliers as missing values and include in the imputation process. Other way is winsorizing, which replaces the extreme values with specific value, such as percentile value or some standard value. However, this process requires great consideration of each feature's property. Sometimes, extreme values can be valuable to catch mysterious symptom. For instance, extremity in some blood serum tests indicates the problems of liver or kidney, which may lose crucial information if we just winsorize or trim the outliers. The other way should be devised for it, one of which is transformation of the feature to hold this information. The way of treating abnormal data should be mentioned in your research paper, which is crucial to prevent misunderstanding of readers.

Next step is about coding and transforming the data, as well as data reduction through diverse ways. Continuous variables need to include nonlinear terms. And categorical variables have to be coded as dummy variables if they are not ordinal. Discretization of continuous variables is not recommended without close attention, because it brings loss of information and overfitting. In addition, score or count measures can integrate several variables into one, such as family history of cardiovascular diseases including separate one. Graph interpretation by plotting univariate relationship between each feature and target value can help to understand how to transform. But it is controversial because this process without confounding also can lead to distortion in statistical inference. Alternative way is to examine the transformation with all other predictors in the model. Data reduction, distinct from variable selection, is the method to reduce the number of independent variables ignoring the target variable. Above mentioned, deleting variables with large missing or narrow/scarce distribution could be covered in this step. Redundancy analysis (RDA) is a method to remove variables that can be easily explained from other variables by pairwise correlation. It could decrease collinearity by ruling out variables relying on others. Principle components analysis (PCA) is another way to produce principle components orthogonally transformed by initial variables, so we can choose few of principle components explaining large portion of original feature's variances to simplify the further model. However, principal component approaches may confuse interpretation of each component and make the components inconsistent with different datasets. Sparse PCA and variable clustering based on PCA can substitute and complement this process, but it requires more advanced techniques to accomplish desired goal. Overall steps are summarized in Table 2.

3. *Model specification and estimation*

Table 2 Summary of preliminary steps with raw data

Steps	Details
Find out and treat abnormal data	Merge category with rare samples or events into adjacent category
	Analyze missing patterns and apply appropriate imputation or truncation methods
	Detect outliers and deals with them by winsorizing or trimming
Code and transform data	Include nonlinear terms, code categorical data as dummy variables
	Discretize continuous variable and make score/count measures only in necessary conditions with careful judgement (not recommended)
	Apply data reduction methods such as RDA, PCA when needed

If potential risk factors are prepared from prior steps, we can get started to make a nice predictive model. Choice of model is the first thing to decide based on target variable. In particular, we focus on survival data frequently appearing in clinical practice, so our target is hazard ratio. Cox proportional hazard model would be mainly employed for our analysis. Nevertheless, most of models share further introduced algorithms for handling model uncertainty.

First concern of model is "how to select important factors appropriately?". If subject knowledge can decide it perfectly, it would be the best. But it is seldom available to choose directly without simultaneous modeling process. Therefore, variable selection during modeling should be reasonable and reliable. Stepwise variable selection is the most widely known and used. This algorithm chooses variables forward or/and eliminate backward by criteria like p-value or AIC (Alkaike Information Criteria). However, it has been largely criticized by many statisticians because the procedures drive into highly biased and overestimated model. And it is known to choose arbitrary variables including noises in simulation studies. Therefore, we need to pay close attention to apply it if necessary. Several statisticians suggested that backward elimination with relaxed criteria (P-value over 0.50) may perform better to deal with collinearity and exaggeration of parameters and statistics.

"How to accurately estimate the coefficients of regression?" is another problem. In numerous cases, the sample size of clinical research is insufficient to fully represent the whole population. For this reason, maximum likelihood method, often used for estimating unbiased coefficients of regression model, could cause testimation bias (overestimation of predictors' effects). Shrinkage is the technique to correct this overestimated effects. There are various ways, for example, simply multiplying heuristic or empirical constants and using resampling method to adjust the coefficients. Currently, penalization method is getting popular after development of ridge regression (L2 penalty) [14]. By adjusting its penalty scale, we can gain required shrinkage. Moreover, LASSO (Least Absolute Shrinkage and Selection Operator, L1 penalty) [15] was proven to reduce dimension of parameters by making insignificant parameters zero. Elastic net [16] is the mixed version of L1 and L2 penalty, which can simultaneously achieve variable selection and parameter shrinkage by (5).

$$L_p(\beta) = L(\beta) + \lambda\left\{\alpha||\beta|| + \frac{1-\alpha}{2}\beta^T\beta\right\} \tag{5}$$

$L(\beta)$ is the log likelihood function of estimated parameter β and $L_p(\beta)$ indicates penalized log likelihood function. λ and α is hyper-parameter for overall magnitude of penalty and the ratio of L1 penalty each. By balancing the hyper-parameters, it is known to take the advantages of both ridge and LASSO about adequate shrinkage and restricting collinearity.

4. *Model evaluation*

Above all, we seek to settle modeling processes for clinical risk prediction. Evaluation of the model should be put in the end of the entire processes together to see whether modeling works well or not. If we have external dataset to validate and generalize the model, it would be great for testing. Before then, we should check the model assumption and internally validate the model performance.

Model assumption is common in any regression. Checking linearity and additivity of predictor should be accompanied with modeling process by considering pre-specified nonlinear terms and interactions. Distributional assumption on specific type of regression also needs to be examined, such as normality and conditional independence of error term in linear regression. Especially, in survival model, time-independence and proportionality of hazard function is fundamental assumption. Additionally, we have to identify whether there exist overly influential observations and why they are so influential. Graphic examination by residual plot is preferred to check these model assumptions, regardless of types of regression because most statistics for checking assumptions are unstable and sensitive in different setting. Usually, models satisfying assumptions tends to prevent overfitting.

For prediction, the model needs to have ability to predict results from new data samples. It is commonly split into two parts, discrimination and calibration. Discrimination indicates how well the model can discriminate the subjects, for instance, high-risk and low-risk patients. And calibration represents how closely the estimated outcomes resemble the actual outcomes. There are many different statistics or visual methods for each model, so we can utilize them in our situation by referring historical models. However, optimism of these measures frequently happens with limited data pool, which seems optimistic but actually poor in new sample because of overfitting. Accordingly, we need to validate the model without apparent illusion. Externally well-defined validation cohort testing the model by other examiners would be the best way for model validation. But the model should go through internal validation enough to convince the model in advance. Cross-validation and resampling methods are usually employed. Among them, bootstrapping is very flexible and extendable methods to make up for any kinds of statistical inference process with incomplete data. It repeatedly draws random samples from original sample to estimate statistics of the model. Thus every procedure mentioned above can be integrated by bootstrapping. Optimism can be calculated by difference between the performance of model with original data (apparent model) and with bootstrap data (bootstrap model) in each sequence. Then optimism-corrected performance can be estimated by subtracting the average optimism from apparent performance. It could be valuable indicator of overfitting. Besides, the specification and estimation of parameters could be strengthened by this nonparametric resampling method so that we can get more empirical distributions from generated samples.

Table 3 Basic information of given random sample

Sex	Sample size	Number of event	Average time to event (Year)
Male	808	29	4.334
Female	1190	32	4.363

3 Case study on Cardiovascular Desease

3.1 Data Description

Around 2000 random samples from published paper [17] was given at the first hand from medical company. The detailed procedure of collecting the data is mentioned in the paper. The clinical trial was implemented for the patients with specific cardiovascular disease in order to track the risk of particular event. Table 3 summarizes basic information of the sample.

As described in the table, it has very limited information about the event, but contains abundant candidates of risk factor. Though, it is frequent situation of medical data in the beginning.

3.2 Procedures Before Experiment

Among more than three hundred features of patients, meaningful factors were chosen by discussing with clinical experts and referring to above literatures of established models in Table 1. Nonlinearity and interactions were also included in the candidate based on clinical view. Rare categories were combined with nearby category or other categorical variables. Discretization of continuous variables was scrutinized with comparison of nonlinear transformation of that variable. Some variables were discretized with clinical standard because they should be treated to detect the abnormality, instead of winsorizing the outliers. All of data preprocessing steps were done with consideration of distributions and correlations. A variety of data reduction algorithms (RDA, Sparse PCA, Variable clustering) were implemented to identify more detailed correlation between variables to remove redundant variables in advance. Patterns of missing values were also looked up for imputation. Because missing data is a small amount, comparison between imputation methods from uniform imputing, to multiple imputation was not significant. To ensure variability of missing values within the dataset, multiple imputation with predictive mean matching (PMM) [18] was applied to fill in the missing except few features that needs clinical intention for imputing. The 10 imputed datasets were generated for subsequent steps. They will be separately used for building models.

3.3 Experiment

The primary objective of the experiment on this small random sample was to address this question, "How can we specify and estimate the risk prediction model properly when given data is not sufficient?". Therefore, two different methods, stepwise selection and Elastic Net, would be compared on the basis of simple Cox regression model with clinical interpretation. In order to draw more realistic distribution of parameters and evaluate overfitting of the models, bootstrapping was embedded in the algorithms with 1000 repeated samplings. Totally 10,000 samples from every multiple imputed data were used to estimate optimism, distributions, frequency of being selected and confidence intervals of coefficients. And the optimization of hyper-parameters (P-value, lambda and alpha) in each algorithm was exhaustively investigated by numerous trials.

3.4 Results

The overall summary of each algorithm is written in Table 4. Performance measure, assisted by optimism correcting, was concordance index (C index [19, 20]), which represents the discrimination of survival data with regard to its ranking of hazard ratios. Calibration measure was not recorded due to scarcity of event data, often leading to extreme gaps. The average number of selected variables in each model was mentioned to compare the performance with respect to simplicity of model, in accordance with the chosen hyper-parameter. We can see that Elastic Net has less optimism even if it includes more risk factors.

Furthermore, the distribution of coefficients selected and estimated from each algorithm were illustrated in Table 5 and Fig. 1 to see their differences, regarding shrinkage effect, variance and biasedness from simple model.

Table 4 Summary of results from candidate models

Model	Variable selection	Degree of freedom[a]	Optimism[b]
Male	Full model	37 (–)	0.147 (0.066–0.258)
	Backward stepwise	15.17 (9–22)	0.134 (0.054–0.230)
	Elastic net	23.46 (17–30)	0.113 (0.043–0.200)
Female	Full model	37 (–)	0.159 (0.075–0.261)
	Backward stepwise	14.22 (9–20)	0.124 (0.049–0.216)
	Elastic net	17.93 (10–26)	0.111 (0.041–0.191)

[a] Mean and 95% confidence interval of degree of freedom from the models with bootstrapping
[b] Median and 95% confidence interval of optimism which was obtained by subtracting test C index applying original data to the model from the apparent C-index directly from bootstrapped samples

Table 5 Summary of major risk factors in candidate models

Risk factors	Full model		Backward stepwise		Elastic net	
	Hazard ratio[a]	Ward χ^2[b]	Hazard ratio[a]	Frequency[c]	Hazard ratio[a]	Frequency[c]
Male						
Age	1.048 (0.955–1.165)	8.866	1.080 (0.961–1.180)	0.5507	1.010 (1.001–1.021)	0.9666
Sbp	1.033 (0.985–1.084)	5.854	1.033 (1.014–1.072)	0.7986	1.004 (1.000–1.009)	0.8721
BMI	0.171 (0.024–1.001)	5.037	0.182 (0.043–0.867)	0.9287	0.979 (0.953–0.998)	0.8985
BMI (square)	1.034 (0.998–1.076)	7.473	1.034 (1.015–1.064)	0.8691	0.9997 (0.999–1.000)	0.8435
Female						
Age	1.017 (0.941–1.098)	2.681	1.056 (0.936–1.117)	0.3208	1.003 (1.000–1.010)	0.5818
Sbp	1.015 (0.981–1.051)	1.027	1.025 (0.976–1.050)	0.4816	1.001 (1.000–1.005)	0.5362
BMI	0.714 (0.294–2.859)	1.855	0.551 (0.290–4.164)	0.4789	0.9934 (0.980–1.000)	0.5860
BMI (square)	1.006 (0.978–1.023)	1.049	1.011 (0.970–1.024)	0.4664	0.9999 (0.9997–1.0000)	0.5310

[a] Median and 95% confidence interval of hazard ratio of each risk factors from bootstrapping

[b] Mean Ward χ^2 of each coefficient from bootstrapping, which imply the importance of the risk factor

[c] Selection frequency of each coefficient by the model with bootstrapping

The results show Full model and Stepwise brings wide distribution of estimated parameters, even going beyond opposite direction. Though these major risk factors are known as important, Stepwise selection looks arbitrary reflected in frequency. And the Stepwise selection seems inflating the coefficients of parameters. While p-value criteria become smaller, the inflation becomes severe. In contrast, Elastic Net resulted in much narrower distribution of parameters and caused shrinkage effect of parameter. Frequency of Elastic Net also properly reflects the importance of risk factor corresponding to Ward χ^2 in Full model. And shrinkage effect and narrow distribution appeared in most case of different settings. And the degree of shrinkage and narrowing increases as the magnitude of L2 norm increases

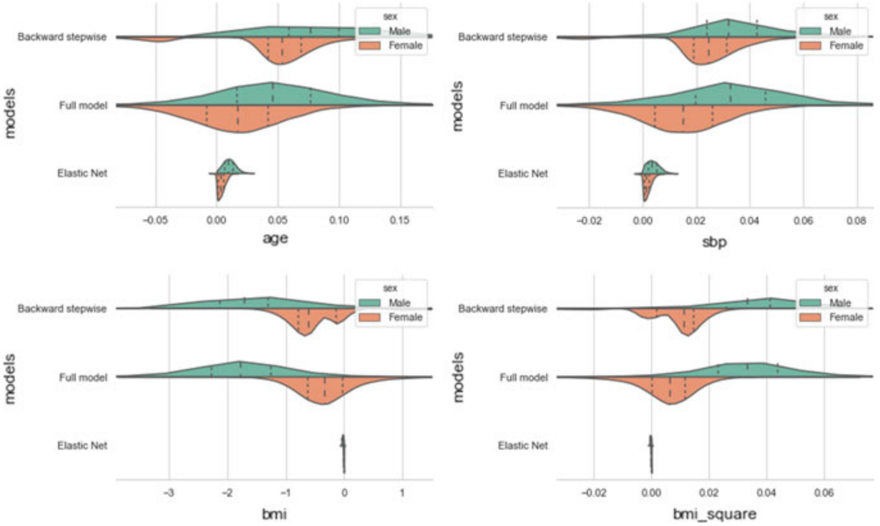


Fig. 1 Bootstrapping distribution of major risk factors in the model

4 Discussion

4.1 Comparison of Variable Selection Techniques

Above experiment show the dramatic difference between variable selection techniques, Stepwise selection and Elastic Net. Stepwise selection is the easiest way to choose variables. However, there have been a lot of concerns from statisticians. As shown in the experiment, the selection of major factors is somewhat at random. This instability of feature selection is one of the severe problems. It includes variables by their estimated coefficients, not true values throughout each “step” causing biased estimation. It is often called as “Winner’s curse”. This biasedness recursively pervades throughout distribution of estimated coefficients, then causes testimation bias and misspecification of statistics. Even noise variables can remain and become significant in the model. As a result, the choice of variables by stepwise can be far from expected list of crucial variables. Also, we can notice that the outcome of stepwise method is overfitted due to testimation bias. Extensive analysis about stepwise selection was summarized in [21]. Previous literature [22] recommended that backward stepwise algorithm with bootstrapping and generous stopping rule may reduce these limitations. But our results showed the problems still remained as mentioned in [23]. It seems that stepwise algorithm not only exaggerated the estimated coefficients, but also broadened the distribution of it. Like female age factor, the distribution of coefficient can even be split into two opposite sign. And frequencies, which means importance of each feature during selections with bootstrapping, were not consistent with other models. It does partly explain arbitrary selection of features. During detailed analysis, testimation bias grows up while the stopping rule become strict. We can infer that more stepwise iteration may cause more "Winner's curse". Therefore, it is found that the optimism of stepwise selection is larger than Elastic Net even if the number of selected factors is the smallest. We can say that unstable choice and testimation error of factors might cause overfitting the model as whole.

On the contrary, penalization method has received spotlight from researchers. Elastic Net, which combined L1 and L2 norms, is widely used in medical research. It can bring shrinkage effect during estimation and shrinkage to zero for variable selection. The penalization could prevent testimation error, which usually occurs when the number of samples is much small compared to the number of variables. As well, it doesn't require sequent ways to select variables, which could cause arbitrary exaggeration by estimated coefficients. Thus it is better way to get adequate candidates of significant factors, as well as to reduce optimism. As shown in Fig. 1, bootstrapped distribution of coefficients from Elastic Net is largely shrunk compared to other two models. It is because of shrinkage effect from penalization and tends to depend on the magnitude of L2 norm. Frequency similarly represents importance of features as Wald χ^2 in Full model. Consequently, the optimism of Elastic Net model was the lowest with reduction of dimension. In sum, penalization technique is superior to stepwise selection by leading to shrinkage of coefficients and reducing bias which can happen during process. Especially in the circumstance that the sample is insufficient, it is one of the best way to decrease overfitting.

In addition, bootstrap sampling should be highlighted while investigating the experiment. Bootstrap model selection [24] is not a novel concept, but its usage has been restricted until nowadays. One of the reason is computational demands. Thus bootstrapping is only used for parameter estimation with efficient algorithms, rather than exhaustive modeling process. However, within finite power of sample, it can exploit the potential power of sample and generate the empirical distribution of any parameter and statistics. And it can calculate and compare optimism of model performance to judge overfitting of a model. This process of modeling can partially cover variation from sampling, making the model robust. Furthermore, in the case of incomplete data, we can deal with variation from missing data by integrating multiple imputation process with resampling. As shown in our experiment, imputed datasets could be augmented by bootstrapping to produce more varied resamples. Detailed descriptions of variable selection under multiple imputation using the bootstrap were mentioned in [25] for prognostic study. This paper raised the problems of stepwise selection procedure using bootstrap as well. Backward stepwise selection using bootstrap sampling is investigated by P. C. Austin [26, 27]. But his conclusion is that the limitation of backward variable selection cannot be solved via bootstrapping method. In contrast, the penalization method using bootstrap sampling has still been developed and verified to devise efficient optimization algorithms and understand its mechanism.

4.2 Limitation and Further Improvement

This paper made two contributions on healthcare services. Firstly, we provided generalized ways to establish clinical risk prediction at the start point with a sparse dataset. It will guide the initial steps for those who wants to enhance medical services with clinical risk management. Second, we gave insight from experiments with modern statistical methods, variable selection and resampling. However, there are several limitations during our trials.

In our experiment, we compared the difference between stepwise selection and penalization method in the situation that small effective sample size is given. We only analyzed about model discrimination performance (C index) and distribution of coefficients. Because the random sample used for our case study includes only few events as Table 3, it is elusive to achieve good calibration. It appears that lack of effective sample size greatly distorts the model specification regardless of any method. In addition, the assumption test was not conducted owing to the same reason. There is another problem of choosing some variables from bootstrapping results to get reduced model for the test. Furthermore, penalization methods require optimizing tuning parameters, in the case of Elastic Net, λ and α. Tuning the parameters is a crucial step to make an adequate predictive model. But there are few available algorithms, even less with Cox's model. Stability selection [28] was employed, but it took enormous time in each trial and brought disparate results in male and female model. Thus we found appropriate tuning parameters with several trials and compared them by optimism.

To complement above problems, we need to do more statistical investigations. Rather than using our dataset, we can simulate data to evaluate and improve variable selection methods. It is apparent that penalization method established superiority in our experiment. But if we can develop optimization algorithm of penalizing and its features with bootstrapping, it could bring much applicability in real situation. As well, we can further analyze how we can adjust extraordinary calibration when data is scarce. For variable selection techniques with bootstrapping, it is also curious, "How can we determine final candidates of risk factors?". It is because selected number of each factor by each method differs. Thus it is another good research question. If we can select the final set of variables, it makes us easy to test the model assumption and validate the model. All these investigations might be well-suited to simulation studies and we can apply the steps to our dataset from simulation results.

Our primary purpose of the modeling strategies was focused on restricted condition when very few data is available at the first hand. Therefore, it cannot attain a perfect solution. Nevertheless, we have learned from the trials, in addition to previously qualified procedure. Following these steps in Sect. 2, from design of model to evaluation, can curtail unnecessary trials and encourage discussion of background knowledge with given data in the beginning of model development. The experiments in Sect. 3 took innovative approaches by investigating the distribution of major parameters and optimism of the models. This experiment showed sharp contrast between the stepwise selection and penalization method in empirical ways. Accordingly, it can give some intuitions for researchers who hesitate over variable selection models.

Acknowledgements In this study, the samples used for case study was provided from AUSA company. Clinical guidance by professionals in the company was of great importance. We particularly appreciate their contribution to our research.

This paper was funded by a grant from the National Natural Science Foundation of China (Grant No. 71971127) and a grant from the Shenzhen Municipal Development and Reform Commission, Shenzhen Environmental Science and New Energy Technology Engineering Laboratory (Grant Number: SDRC [2016]172).

References

1. L. Zhang, H. Wang, Q. Li, M.H. Zhao, Q.M. Zhan, Big data and medical research in China. BMJ **360**, j5910 (2018). https://doi.org/10.1136/bmj.j5910
2. M. Pavlou, et al., How to develop a more accurate risk prediction model when there are few events. BMJ **351**, h3868 (2015). https://doi.org/10.1136/bmj.h3868
3. D. Cox, Regression models and life tables. J. Roy. Stat. Soc. **34**(2), 187–220 (1972)
4. S. Polsterl, P. Gupta, L. Wang, S. Conjeti, A. Katouzian, N. Navab, Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. F1000Res **5**, 2676 (2016). https://doi.org/10.12688/f1000research.8231.3
5. P.A. Wolf, R.B. D'Agostino, A.J. Belanger, W.B. Kannel, Probability of stroke: a risk profile from the framingham study. Stroke **22**(3), 312–318 (1991)
6. C. Dufouil et al., Revised framingham stroke risk profile to reflect temporal trends. Circulation **135**(12), 1145–1159 (2017)
7. J.A. Dorresteijn et al., Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. Heart **99**(12), 866–872 (2013)
8. J. Hippisley-Cox, C. Coupland, P. Brindle, Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. BMJ. **346**, f2573 (2013). https://doi.org/10.1136/bmj.f2573
9. X. Xing, et al., Predicting 10-year and lifetime stroke risk in chinese population. Stroke, p. STROKEAHA119025553 (2019). https://doi.org/10.1161/strokeaha.119.025553
10. STEYERBERG, E.W, Clinical prediction models. A practical approach to development, validation, and updating. J. Roy. Stat. Soc. **66**(2), 661–662 (2010)
11. E. Vittinghoff, D.V. Glidden, S.C. Shiboski, C.E. McCulloch, *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models* (Springer Science & Business Media, 2011)
12. F.E. Harrell Jr, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* (Springer, 2015)
13. J.A. Sterne, et al., Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338, b2393 (2009). https://doi.org/10.1136/bmj.b2393
14. A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
15. R. Tibshirani, Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
16. T. Hastie, H. Zou, Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. **67**(5), 768–768

17. Y. Huo et al., Efficacy of folic acid therapy in primary prevention of stroke among adults with hypertension in China: the CSPPT randomized clinical trial. JAMA **313**(13), 1325–1335 (2015). https://doi.org/10.1001/jama.2015.2274
18. T.P. Morris, I.R. White, P. Royston, Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med. Res. Methodol. **14**(1), 75 (2014)
19. F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, R.A. Rosati, Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
20. M.J. Pencina, R.B. D'Agostino, Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat. Med. **23**(13), 2109–2123 (2004). https://doi.org/10.1002/sim.1802
21. S. Derksen, H.J. Keselman, Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br. J. Math. Stat. Psychol. **45**(2), 265–282 (1992)
22. W. Sauerbrei, M. Schumacher, A bootstrap resampling procedure for model building: application to the Cox regression model. Stat. Med. **11**(16), 2093–2109 (1992)
23. D.G. Altman, P.K. Andersen, Bootstrap investigation of the stability of a cox regression model. Stat. Med. **8**(7), 771–783 (1989)
24. J. Shao, Bootstrap model selection. J. Am. Stat. Assoc. **91**(434), 655–665 (1996)
25. M. W. Heymans, S. van Buuren, D. L. Knol, W. van Mechelen, H. C. de Vet, Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol **7**(33) (2007). https://doi.org/10.1186/1471-2288-7-33
26. P.C. Austin, J.V. Tu, Bootstrap methods for developing predictive models. Am. Stat. **58**(2), 131–137 (2004). https://doi.org/10.1198/0003130043277
27. P. C. Austin, Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. J. Clin. Epidemiol, **61**(10), 1009–17 e1 (2008). https://doi.org/10.1016/j.jclinepi.2007.11.014
28. N. Meinshausen, P. Bühlmann, Stability selection. J. Roy. Stat. Soc. **72**(4), 417–473 (2010)

Pragmatic Interoperability for Extreme Automation and Healthcare Interoperability and Continuity

Jinan Fiaidhi, Sabah Mohammed, and Sami Mohammed

Abstract We are living in a world that is wired differently where almost everything can be made ubiquitously with high degree of connectivity. Such ecosystem of ubiquitous connectivity is a critical piece for the advancements of key industries like manufacturing and healthcare. Both of these industries are transitioning their delivery system from volume based to value based and are looking for new ways to maintain continuity, cut costs, improve quality and increase the levels of interoperability. Accomplishing these goals will help to create a seamless workflow and build a foundation to advance value-based industry. However, manufacturing and healthcare organizations today are balancing the complexities of the using innovative digital transformation technologies with the fusion of working with a multi-disciplinary working teams, designers, developers and customers to achieve enterprise resilience aligned to fiscal and fiduciary responsibility, customer commitments and values, regulatory and compliance requirements and stakeholders' expectations. To integrate such diverse range of perspectives and turn those into meaningful insights requires the planning and enforcement of "Pragmatic Interoperability". This paper describes the authors' vision in developing a flexible workflow infrastructure for enforcing the pragmatic interoperability in industries like manufacturing and healthcare. This vision is based on business continuity planning, web services interoperability, Node-Red, Neo4j and IFTTT workflow technologies.

Keywords Extreme automation · Web services · Services workflow composition · Node-red · IFTTT technologies · Business continuity planning

J. Fiaidhi (✉) · S. Mohammed
Department of Computer Science, Lakehead University, Thunder Bay, Canada
e-mail: jfiaidhi@lakeheadu.ca

S. Mohammed
e-mail: sabah.mohammed@lakeheadu.ca

S. Mohammed
Department of Computer Science, University of Victoria, Victoria, Canada
e-mail: smohamme@uvic.ca

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_3

1 Introduction

Many industries learn the lesson of dealing with the complexity of distributed fusion systems and employ variety of advanced techniques that are more effective and reducing risks [1]. These advanced techniques include technologies like Just-in-time (JIT) systems, Computer integrated manufacturing (CIM), Concurrent engineering (CE), Flexible manufacturing systems (FMS), Collaborative manufacturing (CM), Meaningful Use (MU) and Multisensory Data Fusion (MSDF) to produce products and services with high quality, low costs and shorter lead times [2]. However, enterprise collaboration must involve new type of human engagement, involvement and collaboration as well as proactive planning process that ensures critical services or products are going to be delivered during a disruption. Although the transformation towards such socio technical ecosystem is underway to boost productivity, minimizing risks, simplifying human jobs and ultimately accelerating the industrial growth across geographies, the challenges remain are huge as this integration involves a fusion of technologies that is "blurring the lines between the physical, digital, and human spheres" [3]. It is set to disrupt people, society, business, and government through its innovations. The new face of industry needs to be defined by a confluence between technical complexity and the social system in which the technical complexity is embedded. This new face includes an ecosystem of extreme connectivity, extreme automation, extreme environments, cloud computing delivery, smart digital devices, human and machine collaborations and services interoperability. In the paradigm of manufacturing, this type of ecosystem is focused on using the notion of 'Smart Factory' enriched with capabilities to encourage interoperability and collaboration including [4].

1. Electronic Kanban's instead of Legacy Production Orders
2. Production Surveillance and Prediction
3. Total Productive Maintenance
4. Real Time Data Analysis
5. Virtual Value Stream Mapping.

However, researchers from the paradigm of organizational semiotics (OS) criticized this approach as it is lacking the social level [5–7]. To develop a validated OS research model, a two-stage research design need to be adopted including an assessment study and planning study. The assessment study confirms factors identified from existing experiences as well as to explore any new important factors from empirical and experience context based on sound techniques like:

- Situational Assessment and Risks Analysis
- Business Impact Analysis
- Disaster Recovery Investigation
- Fusion Interoperability Mocking.

The assessment study constructs the industry model based on a top-down black-box approach, which organizes the observations of the operational systems into structures

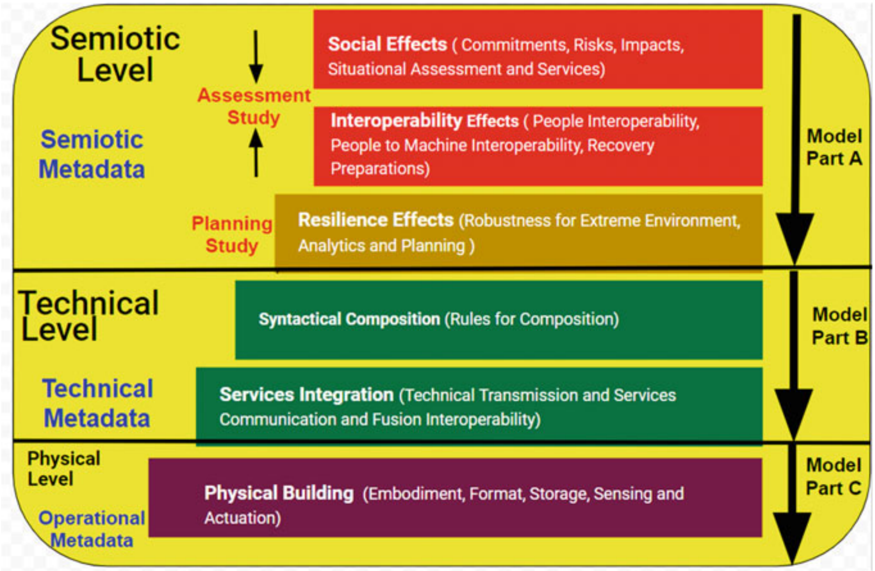

Semiotic Level
Semiotic Metadata
Assessment Study
Planning Study
Social Effects (Commitments, Risks, Impacts, Situational Assessment and Services)
Interoperability Effects (People Interoperability, People to Machine Interoperability, Recovery Preparations)
Resilience Effects (Robustness for Extreme Environment, Analytics and Planning)
Model Part A
Technical Level
Technical Metadata
Syntactical Composition (Rules for Composition)
Services Integration (Technical Transmission and Services Communication and Fusion Interoperability)
Model Part B
Physical Level
Operational Metadata
Physical Building (Embodiment, Format, Storage, Sensing and Actuation)
Model Part C

Fig. 1 A Top-down organizational semiotic framework for designing sociotechnical systems

of agents and affordances interlinked by ontological dependencies and governed by metadata norms. The planning study is to validate the resilience of this model in order to ensure the presentation of a reliable and coherent picture to the stakeholders as well as to fuse data streams and services in real-time. Figure 1 illustrates the overall idea of semiotic paradigm in developing an integral model in top-down way starting from Part A and ending with Part C.

In the semiotic model, metadata information is indispensable for data sharing and interoperability in all the parts of the system model, however, only rarely are metadata available or can be designed in a structured, comprehensive, and machine-readable form. This poses a severe technical challenge for finding and retrieving data as well as for the integrating the various components of the sociotechnical system. The downside of the semiotic "top-down" approach in which authorities defines the "standard," of collaboration which is not flexible and open enough and will inevitably lag behind the ever-changing requirements due to the dynamics in developing sociotechnical systems. Furthermore, this "standard" would need to be accepted by the collaborating parties and components. Several other researchers developed more dynamic models to deal with interoperability and fusion integration. The models eliminate the need for repeated modules across areas of risks, making it easy to expand into new areas when ready, operating as a completely integrated program on a common information foundation. In this direction there are four models that have been taken as a lingua franca for data fusion problems including:

- JDL model [8, 9]
- DDF model [10]

- Omnibus model [11]
- Perceptual Reasoning model [12].

However, none of these models deal with multiple data of different types as well as account for human processing. All these models perform automatic processing of signals from machines and does not account for human interactions. In a typically system architecture one need to support the role and interactions of users as well as fusion of components and services. This paper proposes an extension to JDL Model where it can support different types of data integration and user interactions as intrinsic part of the fusion system for resilience business and continuity.

2 Proposed Methodology

Pragmatic interoperability has been highlighted as a key requirement to enhance collaboration [13]. However, achieving this requirement is not a simple task. The interoperability of enterprises is a challenging goal that could be solved by using notions from different approaches where different enterprises can inter operate smoothly without any particular manual effort and where the dynamism and the autonomy aspects are supported. The Joint Directorate Laboratories (JDL) model addresses this challenge, which was researched in 1993 to deal with multi-level data fusion interoperability and integration, primarily was for military systems [14]. The idea was that you could look at the state of the world considering different entities as “atomic units” of your “World View”. Level 1 Fusion, for example, was called “Object Refinement” and was focused on fusing multiple heterogeneous data sources to obtain information about individual objects (e.g., people, vehicles, buildings, etc.). So-called Higher-Level Fusion dealt with Levels 2–4, which were named “Situation Refinement”, “Threat Refinement”, and “Process Refinement”. Note, there is a big conceptual jump from Level 1 to Level 2. Level 1 consists of real, measurable objects. The other levels are concepts and exist in “linguistic space” instead of physical space. In 1998, the JDL model was upgraded to more general, less militaristic language. There was also a “Level 0” added. So the levels now were “Sub-Object Assessment”, “Object Assessment”, “Situational Assessment”, “Impact Assessment”, and “Process Refinement". Sub-Object Assessment referred to using data to resolve things smaller than individual objects. So it could be an arm, or it could refer to integrating (“fusing”) consecutive reflected pulses from a radar to form a signal [15]. In 2015, a researcher [16], extended the four levels JDL model into five level one where Level 5: “User Refinement” involves humans to be “in the loop” affecting the products of all the lower levels.

However, the increasing amount of information available for planning advocate for the adoption of machine learning methods to address specific situation assessment or risk. A special focus needs to be placed on prognosis and not only diagnosis. With prognosis we can estimate and anticipate events of interest regarding assets and contributing processes. Prognosis allows us to learn about future events while diagnosis learns from the past to predict basic similar events. Machine learning is the core challenge of business modeling and planning as data-driven prognostic approaches aim at predicting when an abnormal behavior is likely to arise within the monitored process, providing further insights such as its severity and impact on the business performance. Thus, it becomes particularly interesting to characterize normality properly towards unveiling degradation patterns or trends. For this reason, adding a sixth level to JDL to include thick data and machine learning analytics is one of the most important direction for business modeling and planning. The idea of having the sixth level is to characterize the business values (including user values) through qualitative measures (aka Thick Data [17]) as well as to identify behavioral patterns of interest on the basis of the data monitored from the process or asset under study (training data) by means of quantitative analytics and machine learning models. This acquired knowledge (valued and thick data as well as quantitative indicators) can be then used to tackle a wide variety of planning problems, including focused prediction, classification and anomaly detection, among others. Furthermore, since data and services are the most important asset that a business owns, and in today's climate of data security and stewardship it's more important than ever to trace how that data was produced and the journey it has undertaken through the workflow of services leading to its present state. What we need in this direction is a strong data and services governance platform. By adding a sixth level to JDL to include machine learning analytics is one of the most important directions for accommodating a risk-centred approach to the planning process. Figure 2 illustrates our interoperable bedside care planning framework (BPF) that can be used for predicting outcomes of several of the frequent emergency department cases as defined by the chief complaints to be the grouped into eleven categories: respiratory, gastrointestinal (GI), undifferentiated infection (UDI), influenza-like illness (ILI), lymphatic, skin, neurological, pain, dental, alcohol and musculoskeletal syndromes. This framework has been proposed by the second author as the new fail-operate architecture for healthcare enterprises to deal with serving large scale COVID-19 cases.

A number of technologies (e.g. Web Services and Agent technologies) have proposed the use of JDL data fusion model to enhance the capabilities of the network and aid in the development of situational awareness for the enforcement of business continuity, interoperability and scalability. In one hand, the Web service technology is used to support interoperability but failed to support the dynamism and the autonomy aspects. On the other hand, agent-based modeling technology is an emerging field that can help understand business dynamics where a set of agents is used to represent various actors in an environment. However, hurdles still exist in agent technology universal adoption because of the diversity of agent communication standards. For example, solving agent interoperability from a protocol and connectivity perspective is straightforward: Either using TCP/IP and/or Ethernet. However, when we peel back the layers of agent system interoperability, we eventually reach a roadblock in the form of software. How do developers get the software stack from a programmable logic controller (PLC) or programmable automation controller (PAC) from vendor A to communicate with the predictive analytics software from vendor B? For example,

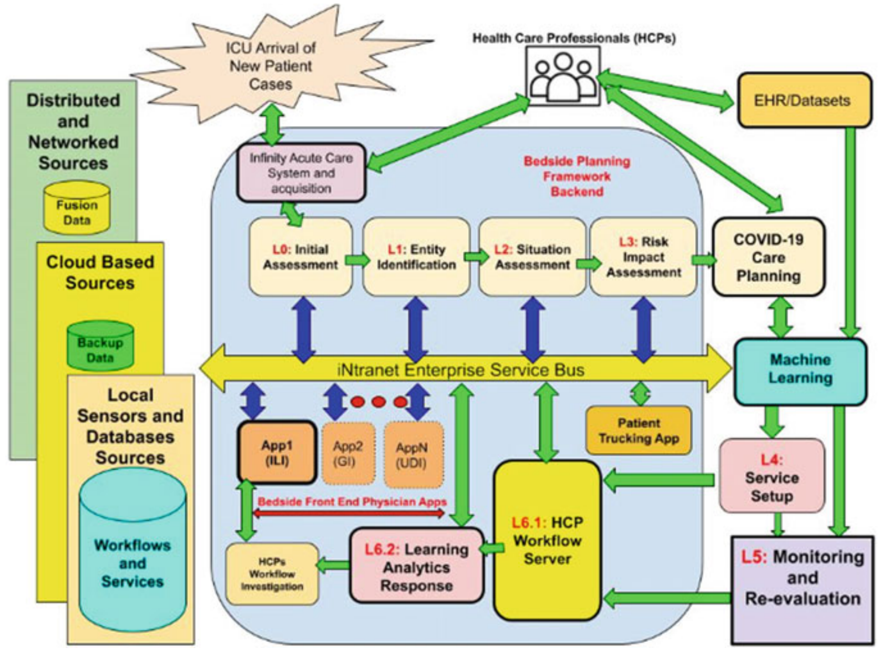

ICU Arrival of New Patient Cases
Health Care Professionals (HCPs)
EHR/Datasets
Distributed and Networked Sources
Fusion Data
Cloud Based Sources
Backup Data
Local Sensors and Databases Sources
Workflows and Services
Infinity Acute Care System and acquisition
Bedside Planning Framework Backend
L0: Initial Assessment
L1: Entity Identification
L2: Situation Assessment
L3: Risk Impact Assessment
COVID-19 Care Planning
iNtranet Enterprise Service Bus
Machine Learning
App1 (ILI)
App2 (GI)
AppN (UDI)
Patient Trucking App
Bedside Front End Physician Apps
L4: Service Setup
L6.1: HCP Workflow Server
L6.2: Learning Analytics Response
HCPs Workflow Investigation
L5: Monitoring and Re-evaluation

Fig. 2 The extended JDL model for healthcare enterprise fail-operate system

how would a developer move data from an automation controller into a cognitive analytics application like the IBM Watson IoT platform? More importantly, how does the developer do that without countless development cycles or hours of manpower configuring and troubleshooting complex middleware? The objective of extreme automation and healthcare interoperability is to connect sensors, industrial devices, computing systems and online services and applications together to provide plant and enterprise personnel with actionable information and cut costs. Today's such environments are changing, evolving, expanding. One need a single source for integrated solution tailored to the business specific workflow automation needs. The solution needs to enforce several services like:

- Add new automation workflow across multiple the business ecosystem
- Accelerate production and improve communications
- Automate workflows
- Automate third-party.

Workflows capture the state of a business process and enable state transitions when a trigger is received. This will enable disparate devices to be into connected and to automate the components of the workflow. Hence adopting a workflow methodology provides a robust means of describing applications consisting of control and data dependencies along with the logical reasoning necessary for distributed execution.

Designed to simplify extreme automation development, many workflow sandboxes deliver a complement of software components needed to move data from end nodes to the cloud and deliver information to end users. Workflows offer developers a simple but flexible programming model at a level of abstraction closer to the domain-specific activities that they seek to perform. However, languages for describing workflows tend to be highly complex, or specialized towards a particular domain, or both. However, Node-Red provided more general and flexible workflow programming paradigm which is originally designed for extreme automation [18]. This paper presents ExtremeFlow: an approach to workflow programming which combines the advantages of a declarative workflow description and web services programming. The workflow execution model of ExtremeFlow is based on a formal model of computation over web services. The execution environment is implemented on the basis of a widely adopted runtime platform node.js.

3 The ExtremeFlow Approach

Web service platforms become the Trojan horse to enforce interoperability with simple protocols like NodeJS. Each web service platform exhibits the following characteristics [19]:

- Each web service encapsulates behavior with low coupling
- Web Services interact with each other using NodeJS over HTTP
- Web Services can be discovered at run-time
- Web Services can be orchestrated to perform a series of functions in a workflow.

With Node-Red web services, healthcare workers can produce fully customizable workflows with flexible connector operators such as loops, data storage, array mapping, branching, and if/then conditionals and many more. One can use his or her mobile device to hook to such services via creating a bucket through webhook relay and installing node-red on the mobile phone (e.g. Android). Figure 3 illustrates our overall microservice to classify new cases arriving at an emergency department as candidate for SEPSIS Shock or not.

However, composing web services as the one described in Fig. 3 into a comprehensive application can be a tedious and error prone task when using traditional textual scripting languages like WSDL [20]. As an alternative, complex interactions patterns and data exchanges between different Web services can be effectively modeled using a visual language. In this paper we are presenting the design of the ExtremeFlow composition ecosystem. ExtremeFlow uses visual composition that has been fully implemented in a development environment for Web service composition with usability features emphasizing rapid development and visual scalability. The ExtremeFlow environment provides an integrated toolkit to manage the whole lifecycle of a workflow composition in any business (Fig. 4).

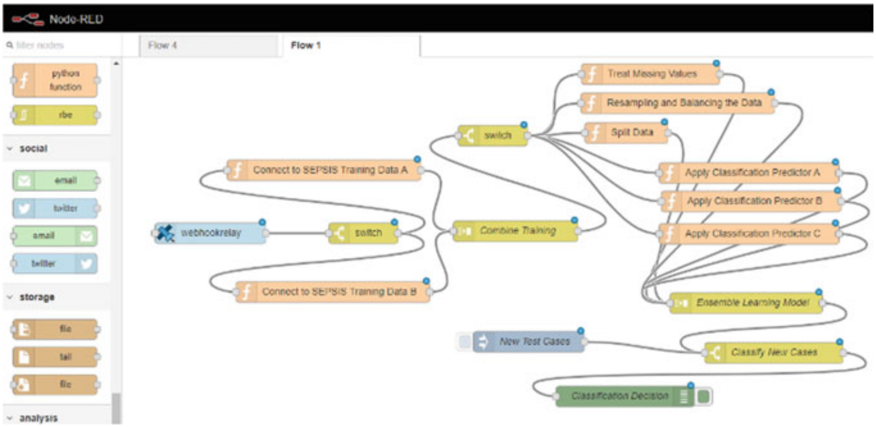


Fig. 3 NODE-RED web service workflow for predicting septic shocks for COVID-19 cases

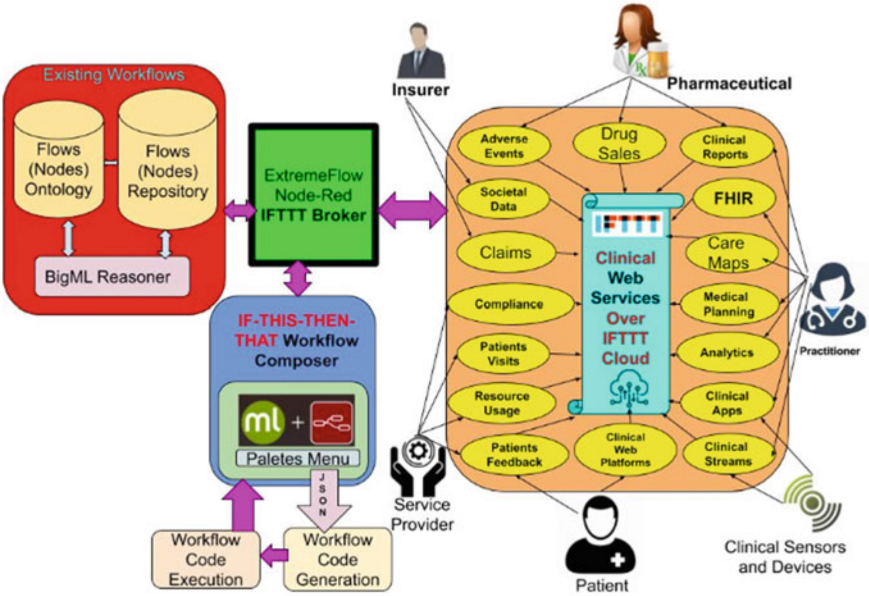
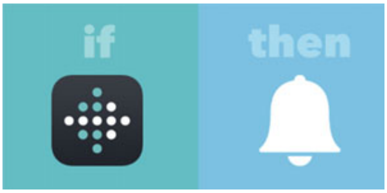


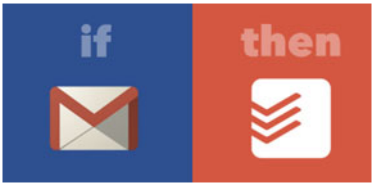
Fig. 4 Implementing the ExtremeFlow using the IFTTT Broker for eHealth workflow applications

In order to manage and fuse these services, an architecture is required that integrates the services in a loosely coupled way to support decentralized discovery and workflow execution environment. Enforcing this flexibility, the ExtremeFlow have included the IFTTT broker. This broker provides library of methods to allow services to send and receive messages and data through the workflow chain. The ExtremeFlow provides a simple mechanism to compose workflows using a rule structure of Triggers and Actions (called rule applet). An applet is triggered by changes that occur within other web services such as Facebook, Telegram, Instagram, or Pinterest. Figure 4 illustrate the usage of the Node-Red IFTTT broker in developing and composing workflows in the healthcare paradigm.

The composition process using the IFTTT broker begins with the Flow Repository, where Web services can be imported as reusable components. Any user can search the repository with the help of ontology and a reasoner, select a set of existing services and drag them into the workflow composer. Then, the new workflow can be modified by adding new nodes from the palate’s menu or from the web services published over the cloud. This operation is partially automatic, since the editor can bind parameters with matching names. To get an overview over the order of execution of the tasks and add additional constraints, the user may view and edit the workflow anytime. Once all of the services have been connected the new process may be compiled and uploaded to an ExtremeFlow runtime environment for execution. The IFTTT workflow composer uses a flexible domain specific language (DSL) to solve a range of problems in the eHealth application domain [21, 22]. There has been a range of DSLs (e.g. Tasker, IFTTT) for heterogeneous systems like those for IoT, to enhance the programmability or simplify the process of wiring different components of an IoT system [23]. This type of DSLs allows users to create workflows with “triggers” and “actions”. The “IF THIS” keyword is used to identify the triggers and “THEN THAT” is to identify the actions part. The trigger is activated by changes that occur within other web services such as FITBIT steps or an email from Gmail. Upon activation you may assign an action like pushing a reminder notification about your achieved steps:



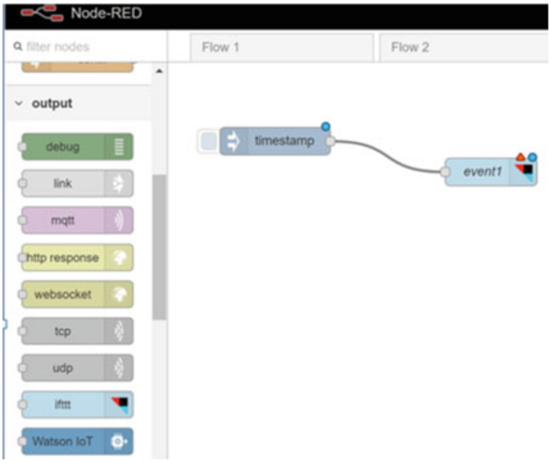
and you may chain this part of the workflow with another part like if a notification from Gmail arrived then you will add it to your TODO list:



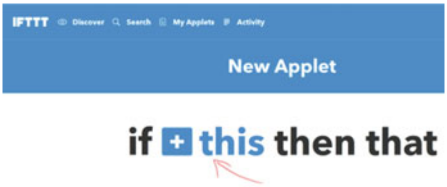
The automations are accomplished via Node-Red recipe with IFTTT[1] integration—which are sort of like macros that connect multiple apps and devices to run

[1] https://ifttt.com/

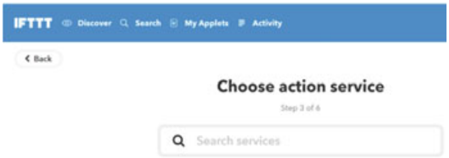
automated tasks. As an example, to illustrate using our ExtremeFlow composer to detect the presence of patients wearing a Fitbit Flex[2] near a Bluetooth (BLE) scanner.[3] In the composer we will need to create an initial flow where we will drag one IFTTT node (now called event1):



However, we will need to define by clicking on the IFTTT event 1 where we will be directed to https://ifttt.com/my_applets to create a new IFTTT applet:



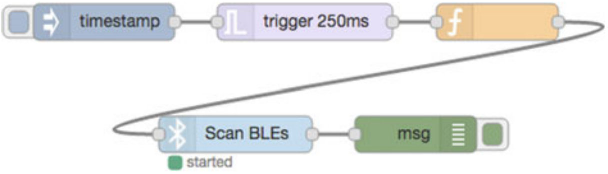
By choosing the BLE Scanner this trigger will be set to this service and we will need to set the action required when the trigger will be activated.



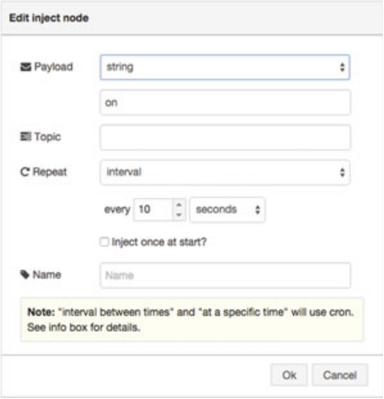
[2] https://www.fitbit.com/en-ca/flex2

[3] https://developers.sensetecnic.com/article/detect-presence-using-a-fitbit-node-red-and-fred/

Now a complete workflow will look like this:



We can inject a message to scan for the Fitbit Bluetooth availability every 10 s:



Edit inject node

Payload: string

on

Topic:

Repeat: interval

every 10 seconds

□ Inject once at start?

Name: Name

Note: "interval between times" and "at a specific time" will use cron. See info box for details.

Once this is done, bring a Fitbit flex in range to the computer and deploy the flow. If all is well you should should see output similar to Listing 1 in the debug pane when it discovers devices.

{ "payload": { "peripheralUuid": "88e1ab7bcfa4493dbaf16c55e69a4881", "localName": "Flex"}, "peripheralUuid": "88e1ab7bcfa4493dbaf16c55e69a4881", "localName": "Flex", "detectedAt": 1,452,023,547,186, "detectedBy": "myhost.local", "advertisement": { "localName": "Flex", "txPowerLevel": -6, "serviceData": [{ "uuid": "180a", "data": [4, 7]}], "serviceUuids": ["adab1fb86e7d4601bda2bffaa68956ba"]}, "rssi": -89, "_msgid": "a017136d.5fe8f"}.

4 Conclusions

We tested our ExtremeFlow composer on a Google Pixel 3 running Android 9. In every application domain, interoperability between various services comprises the integration of computation, software, networking, and physical processes. Consequently, interoperability models required an increasing support for hybrid and heterogeneous models, networking, services and time synchronization. To assist developers and users alike in designing such systems, we have developed a prototype for composing workflows based on incorporating an IFTTT Broker (IF-THIS-THEN-THAT) Domain-Specific Language (DSL) that comes with fully-automated Node-Red and IFTTT tool support. Our ExtremeFlow composer includes support for: (i) interactive Node-Red model description with input validation; (ii) the computation of possible operation modes of subsystems and parts; and, (iii) checking the adherence to requirements for various design alternatives and finding the near optimal designs given these requirements. Moreover, the generated workflow models provide visualizations throughout the toolchain which help design engineers to better understand the implications of design decisions and communicate them to stakeholders. The ExtremeFlow composer has been applied to the healthcare domain. We are currently experimenting to extend the ExtremeComposer to provide smart analytic capabilities through integrating the Node-Red with the BigML services.

Acknowledgements This research is supported by the author NSERC DDG 2020. We would like also to thank the LISS 2020 Conference Organizers for accepting this paper (July 25-28, 2020 Budapest, Hungary).

References

1. M.E. Liggins, C.Y. Chong, I. Kadar et al., Distributed fusion architecture and algorithms for target tracking. *Proceedings of IEEE*. pp. 95–107 (1997)
2. B. Zhao, F. Steier, Effective computer integrated manufacturing (CIM) implementation using socio-technical principles. Ind. Manage. **35**(3), 27 (1993)
3. K. Schwab, The fourth industrial revolution: what it means, how to respond, World Economiv Forum, 14 Jan 2016, Available Online: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/
4. R. Davies, T. Coole, A. Smith, Review of socio-technical considerations to ensure successful implementation of Industry 4.0. Proc Manufact **11**, 1288–1295 (2017)
5. K Liu, *Semiotics in Information Systems Engineering* (Cambridge University Press, 2000)
6. J. Cordeiro, J., Filipe,The semiotic pentagram framework—a perspective on the use of semiotics within organizational semiotics, in *Proceedings of the 7th International Workshop on Organisational Semiotics* (2004)
7. K. Liu, Requirements reengineering from legacy information systems using semiotic techniques. Syst. Signs. Actions Int. J. Commun. Inf. Technol. Work **1**(1), 38–61 (2005)
8. A. Steinberg, C. Bowman, F. White, Revisions to the JDL data fusion model, in *Proceeding SPIE*, ed .by Dasarathy, B. V., vol. 3719, pp. 430–441 (1999). Sensor Fusion: Architectures, Algorithms, and Applications III
9. J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, F. White, Revisiting the JDL data fusion model II. December 2004. Available Online: https://pdfs.semanticscholar.org/b183/008f9d63252bd4cc4438cf0bf744bc8e995d.pdf
10. B. Dasarathy, Decision fusion strategies in multisensor environments . IEEE Trans. Syst. Man Cybernet. **21,** 1140–1154 (1991)
11. M. Bedworth, J. Obrien, The omnibus model: a new model of data fusion? AES Magaz (2000). Available Online: https://pdfs.semanticscholar.org/7595/1a01b0c603774ed99e57d9b427ce55741020.pdf
12. I. Kadar, Perceptual reasoning in adaptive fusion processing. SPIE (2002). Available Online: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4729/0000/Perceptual-reasoning-in-adaptive-fusion-processing/10.1117/12.477619.short

13. C.H. Asuncion, M.J. Van Sinderen, Pragmatic interoperability: a systematic review of published definitions, in *IFIP International Conference on Enterprise Architecture, Integration and Interoperability* (Springer, Berlin, Heidelberg, 2010), pp. 164–175
14. M.E. Liggins, D.L. Hall, J. Llinas, *Multisensor Data Fusion*, Second Edition: Theory and Practice (Multisensor Data Fusion). CRC (2008). ISBN 978-1-4200-5308-1
15. M. Meloon, What is an example of data fusion? How does it help in big data? Is it data analysis technique? Sep 19, 2014, Available Online: https://www.quora.com/What-is-an-example-of-data-fusion-How-does-it-help-in-big-data-Is-it-data-analysis-technique
16. E. Blasch, One decade of the data fusion information group (DFIG) model, Proceedings Volume 9499, Next-Generation Analyst III; 94990L (2015). https://doi.org/10.1117/12.2176934. Event: SPIE Sensing Technology + Applications, 2015, Baltimore, Maryland, United States
17. J. Fiaidhi, S. Mohammed, Thick data: a new qualitative analytics for identifying customer insights. IT Prof **21**(3), 4–13 (2019)
18. C. Simpkin, I. Taylor, D. Harborne, G. Bent, A. Preece, R.K. Ganti, Dynamic distributed orchestration of node-RED IOT workflows using a vector symbolic architecture, in *2018 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, pp. 52–63. IEEE (2018)
19. D. Ganesarajah, E. Lupu, Workflow-based composition of web-services: a business model or a programming paradigm? in *Proceedings of Sixth International Enterprise Distributed Object Computing* (IEEE, 2002), pp. 273–284
20. C. Pautasso, G. Alonso, Visual composition of web services, in *IEEE Symposium on Human Centric Computing Languages and Environments, 2003. Proceedings, 2003* (IEEE, 2003), pp. 92–99
21. F. van den Berg, V. Garousi, B. Tekinerdogan, B.R. Haverkort, Designing cyber-physical systems with aDSL: a domain-specific language and tool support, in *2018 13th Annual Conference on System of Systems Engineering (SoSE)* (IEEE, 2018), pp. 225–232
22. T. Nägele, J. Hooman, Rapid construction of co-simulations of cyber-physical systems in HLA using a DSL, In *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (IEEE, 2017), pp. 247–251
23. Quirk, C., Mooney, R., Galley, M. Language to code: learning semantic parsers for if-this-then-that recipes, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (2015), pp. 878–888

AHP-Based Multicriterial Ranking Model for the City Logistics Analysis of Urban Areas

Dávid Lajos Sárdi and Krisztián Bóna

Abstract Nowadays, the urban freight traffic is a very important research area because of its significant effect on the urban sustainability. In this field, we can highlight the problems of the logistics systems of the so-called urban concentrated sets of delivery locations, as we need to serve lots of stores in a very small area in case of these sets. This research field is in the focus of the City Logistics Research Group of the Budapest University of Technology and Economics, since 2015. Based on the results of former research projects, we can say, that it could be really important to develop such a ranking model, which could rank the current city logistics state and the future city logistics development potentials of concentrated sets of delivery locations and other urban zones with individual delivery locations in the same time. In our paper, we are going to present an Analytic Hierarchy Process based multicriterial ranking model, which makes it possible to rank urban areas in a two-dimensional system, where the first dimension is the current city logistics state, and the second one is the future development potential. The main purpose of this ranking is to mark in future city logistics projects those urban zones, they could be well developed, and it is worthwhile to develop them.

Keywords Logistics · City logistics · Budapest · AHP · Ranking model · Simulation · Shopping mall

1 Introduction

In the City Logistics Research Group of the Department of Material Handling and Logistics Systems at the Budapest University of Technology and Economics, we started in 2015 our research project to examine the city logistics systems of the

D. L. Sárdi (✉) · K. Bóna
Department of Material Handling and Logistics Systems, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: david.sardi@logisztika.bme.hu

K. Bóna
e-mail: krisztian.bona@logisztika.bme.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_4

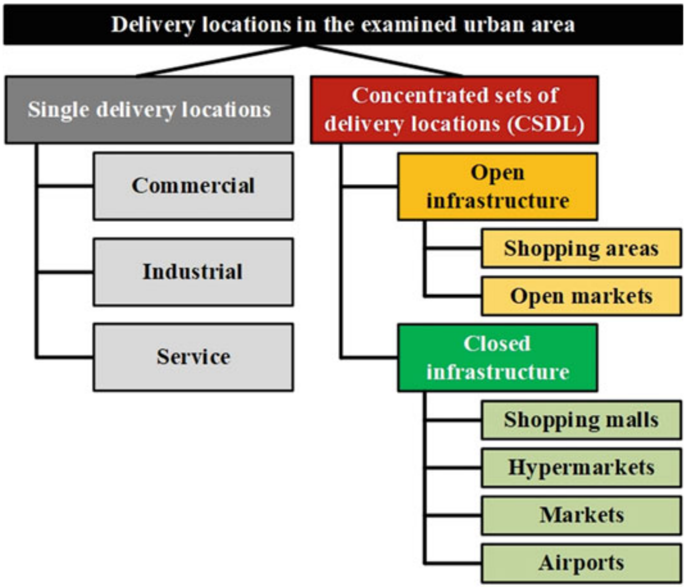


Fig. 1 Groups of the examined delivery locations [2]

so-called urban concentrated sets of delivery locations. For this research, the urban delivery locations can be assigned to two groups (this kind of classification is also a very important result of our research, as the urban delivery locations were not classified in this way earlier): there are single delivery locations and concentrated sets of delivery locations (referring to CSDL; basically, groups of single delivery locations) [1], they are in the focus of our research. As it can be seen on Fig. 1, in case of the examined CSDLs we can define two type of concentration: concentration with open and with closed infrastructure. In case of open infrastructure, the set is defined by an open area, so roads and squares can mark the borders of the set, like in case of a shopping area or an open market. In case of closed infrastructure, the set is defined by a closed area, which means in our case a building, like in case of shopping malls or hypermarkets [1]. Despite the fact that the CSDLs have a very significant goods traffic with significant development potentials based on former city logistics projects, they are not in the focus of logistics researches and at the planning phase of the CSDLs the logistics aspects are not as important as the customer aspects [2]. As a result of this, our Research Group started to examine these CSDLs, we collected data about their characteristics, developed new system concepts and several models [2].

In our paper, we would like to present such a multicriterial ranking model, which can rank the current city logistics state and the future city logistics development potential of urban zones at the same time, but not only for these CSDLs, but for other urban zones as well, which contain single delivery locations. In this ranking model, we would like to examine every relevant logistics properties for the current system

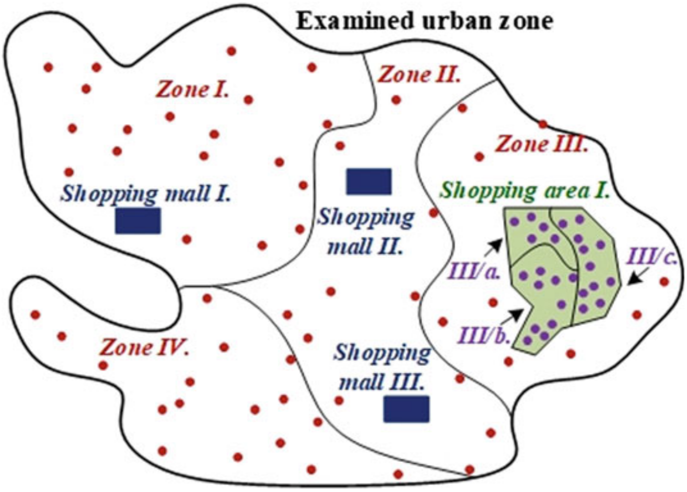


Fig. 2 Example of an examined urban zone with 10 zones

and for the possible future solutions too. As an example, on Fig. 2 an examined urban area can be seen with 10 zones.

In this area, there are 4 simple urban zones with single delivery locations, 3 shopping malls and 1 shopping area (which can be divided to 3 subzones). For this area, our ranking model will be able to decide, which of the 10 zones is currently in the best city logistics state and which one of them could be the best developed in the future. Later, on Fig. 3, we are going to use the same zones.

As a first step of our research, we performed a literature research, where we were looking for multicriterial models in city logistics and in other logistics fields.

2 Multicriterial Models in City Logistics

In the literature review phase, we worked from 3 scientific databases: from ResearchGate, from ScienceDirect and from Google Scholar.

We would like to highlight 3 of the examined ranking models. The first of them deals with bigger networks and its main purpose is to rank logistics centers [3]. A similar model could be implemented for the junctions and even for the zones of a city logistics network. One of the two other ranking models ranks city logistics projects [4], and the other one examines the operation suitability of electromobility concepts for last mile delivery tasks [5], so they have similar aspects as we need for the evaluation of the city logistics development potentials. In case of some other publications, different multicriterial models are used for the evaluation of city logistics developments [6, 7] and for the selection of the best concepts [8]. Above these, for several different purposes can the multicriterial methods be used in city logistics:

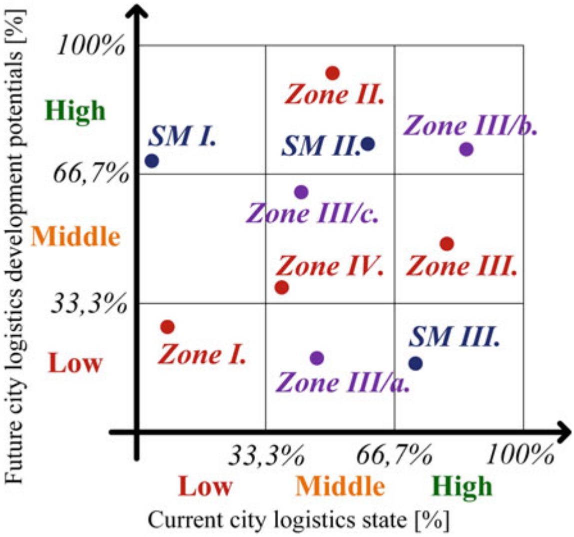


Fig. 3 Example of the results of the two-dimensional AHP-based ranking

for the evaluation of the performance of city logistics terminals [9], for urban sustainability aspects [10], for regulation evaluation [11], for vehicle selection [12] and for urban facility location problem solution as well [13, 14], but our two purposes were not examined together in these researches. In the literature research, we were also looking for the possible multicriterial models, they can be used for our research. For this, we examined more than 50 papers, books and thesis about city logistics projects and from other logistics fields as well. The main results of the literature research can be seen below:

- there are several multicriterial ranking models in city logistics, but none of them has the same purposes as our research, so the ranking model, which examines the current city logistics state and the future development potentials at the same time, will be a new application;
- the Analytic Hierarchy Process method [15] has significantly more documented application in city logistics and in other logistics field, than any other method (more than 50% of all examined papers works with AHP), so we can conclude that AHP is an appropriate method for the evaluation of logistics systems, we are going to use this method.

3 Results of Our Former Research

In this chapter, we would like to highlight our main research results (which project focused on the CSDLs), as they are going to be very important input for our ranking model.

In the first phase of our research, we had to collect data about the stores of the examined CSDLs (from Budapest; as we live and work there, we could focus on this city as a first step), as there was no available data about the logistics characteristics about them or about single delivery locations of any other urban zone. Between 2015 September and 2019 May, we examined 5 CSDLs from Budapest (4 shopping malls and 1 shopping area) and sum 540 stores of them. The data collection was performed by use of our own methodology, which has a complex questionnaire with 31 questions (or 32 in case of shopping areas) about all important logistics parameters (e.g. general properties, vehicles, logistics units, delivery times, e-commerce) as its most important component [16]. Based on our results, in the current city logistics system of the CSDLs, the deliveries are not synchronized with each other, and this means that the big amount of goods is delivered in several parts, in smaller vehicles with low utilization. Based on the data collection, we have daily 0.98 delivery per store. In these deliveries, daily sum 3900 pcs of boxes, 250 pallet unit loads, 3000 clothes hanger units and 1300 other units are handled in the examined stores. This means more, than 80 tons of goods daily, so yearly sum 30,000 tons of goods, only for the responder stores from 5 CSDL from Budapest. As it can be seen, we handle a huge volume of goods, and we could improve the delivery processes by organization solutions [17, 18].

The results of the data collection clearly showed us, that the problems, we expected are real in the examined city logistics system. Next step of our research was to model the current city logistics system of the CSDLs and some new concepts as well [2]. First, we developed the mathematical model of the current and new system [1, 17]. Next, we developed the mesoscopic, MS Excel based simulation model of the current and the new system, first between the city borders and the logistics areas of the CSDLs [18] and then for the whole logistics process between the suppliers and the customers of the CSDLs [17]. Additionally, we developed the mathematical model of the cost structure [19], the topological model of the CSDLs with open infrastructure [17], and the geometrical model based macroscopic simulation model of a special cargo bike city logistics system [20]. In the simulation, we examined 217 stores of 4 shopping malls, we are going to test our AHP-based ranking model for these malls.

Based on the simulation results, in the examined city logistics system in one month, approximately 1750–1781 tons of goods are handled. In the current system, this generates monthly 3990 delivery transactions for the 217 examined stores with additional empties handling as well. For these deliveries, the sum monthly mileage is 780,022 km, which generates for example 137.4 t/month CO_2-emission. To compare with the current system, we examined two new city logistics concepts, where we placed a consolidation center and cross docks to the system, and the consolidated deliveries were handled by bigger road vehicles or by cargo trams [1, 2, 17]. This examined amount of goods generated between the suppliers and the center monthly 1507 delivery transactions; and additionally, between the center and the malls by use of road vehicles 401 consolidated transactions, and by use of cargo trams 174 consolidated transactions. To handle these amounts, the sum mileage in case of road vehicles was 376,736 km (51.7% reduction), and in case of cargo trams it was reduced to 371,963 km (52.3%). In the first case, the sum CO_2-emission is reduced to 55.5 t, in case of cargo trams we can reduce the direct emissions even more.

We also evaluated the logistics costs in both cases, by use of our simulation model [17, 19]. Based on the simulation runs, in the current system for 217 stores of 4 shopping malls, the monthly sum logistics cost is € 450,666 (so nearly half million Euro). In the new system, we can reduce this amount by 38.2% in case of road vehicles, and by 44.3% in case of trams. This means in the first case sum € 270,556 monthly logistics cost and € 250,839 in the second case [20]. This means, that in the new logistics system not only the performances and the emissions can be reduced, but also the costs of the daily operation.

4 Development of the AHP-Based Ranking Model

How to examine the current city logistics state and the future city logistics development potentials at the same time, this was the first important question in the development phase of the ranking model. We expected, that here will be some criteria, they will affect both (e.g. the current state of the infrastructure or the current regulations) and they can have different weights in the cases, we decided to use a two-dimensional solution, where we use two AHP-based model at the same time. One of them will examine the current state and the other one will examine the future development potentials. So, our ranking model will be able to handle both purposes at the same time, and its result will show us, how developed is our urban zone currently and how can it be developed in the future. In this way, the ranking and the classification is realized using by two independent performance indicators. The expected results of the ranking can be seen on Fig. 3 (with the same zones, as we had earlier on Fig. 2), where the points show us the ranking values of the zones, and the zones are assigned to two groups. In our example, the upper and lower limit values of the "high", "middle" and "low" categories are evenly distributed on the 0…100% range. Some typical expected results of the evaluation can be seen on this figure well: there are zones, they are not well-developed, but they can be developed well in the future, e.g. SM I.; there are zones, they are well-developed, but they can't be developed well in the future, e.g. SM III.; there are zones, they are not well-developed, and they can't be developed well in the future, e.g. Zone I.; and there are zones, they are well-developed, and they can be developed well in the future, e.g. Zone III/b.

In the development process, first we formalized the two-dimensional AHP-based model, then we defined the criteria for both dimensions, and finally, the criteria were ranked by experts of the City Logistics Research Group and by some other experts.

4.1 AHP-Based Ranking Model

The first step of the AHP-based ranking is to prepare the comparison matrices for the evaluation of the current state and the development potentials ($\underline{\underline{A^P}}$ and $\underline{\underline{A^F}}$ matrices, where P means Present and F means Future). In the matrices, A_i^P, and A_k^F are the examined criteria. In this paper, we would like to present the formulas of the AHP only for the current city logistics state, in case of the future development potentials, only the indexes are different. The comparison matrix for the current city logistics state can be seen in Table 1, where the w_i^P values are the weights of the criteria, i and j are integer indexes ($i, j = 1 \ldots n$; in case of the development potential, we worked with $k, l = 1 \ldots m$ integer indexes).

Based on this, we can see that the $a_{i,j}^P$ items of the matrix (they show us, how more important is criterion "i" than criterion "j") can be calculated as in (1). In this case, under and over the main diagonal the values are the reciprocals of each other.

$$a_{i,j}^P = \frac{w_i^P}{w_j^P} \tag{1}$$

The next step of the formalization is to calculate the normalized comparison matrices ($\underline{\underline{A^{*P}}}$ and $\underline{\underline{A^{*F}}}$ matrices), where first, we need to sum the columns of the comparison matrix and then every item should be divided by this sum, as in (2). For these new matrices, it is always true, that the sum of the items in one column is 1.

$$a_{i,j}^{*P} = \frac{a_{i,j}^P}{\sum_{i=1}^n a_{i,j}^P} \tag{2}$$

Next step is to calculate the weights of the criteria in the AHP model (W_i^P and W_k^F) values. The main purpose of this is that the sum of the weights must be 1. We can calculate the weights from the formula in (3), where $\lambda_{\max}^P$ is the biggest eigenvalue, $\underline{\underline{E_n}}$ is the $n \times n$ identity matrix, $\underline{W^P}$ is the eigenvector with the weights. From this system of linear equations, we can get the W_i^P and W_k^F weights.

Table 1 The comparison matrix for the evaluation of the current city logistics state

$\underline{\underline{A^P}}$	A_1^P	A_2^P	…	A_j^P	…	A_n^P
A_1^P	w_1^P/w_1^P	w_1^P/w_2^P	…	w_1^P/w_j^P	…	w_1^P/w_n^P
…	…	…	…	…	…	…
A_i^P	w_i^P/w_1^P	w_i^P/w_2^P	…	w_i^P/w_j^P	…	w_i^P/w_n^P
…	…	…	…	…	…	…
A_n^P	w_n^P/w_1^P	w_n^P/w_2^P	…	w_n^P/w_j^P	…	w_n^P/w_n^P

$$\left(\underline{\underline{A^{*P}}} - \lambda_{\max}^{P} \times \underline{\underline{E_n}}\right) \times \underline{W^P} = 0 \tag{3}$$

Next step of the AHP-based ranking is to check the consistency of the matrices. For this, first we need to calculate the P values, from multiplying the row vectors of the original matrices ($\underline{a_i^P}$ and $\underline{a_k^F}$) by the column vectors with the weights ($\underline{W^P}$ and $\underline{W^F}$) and then dividing them by the weights, as in (4). By use of these values, we can calculate the consistency index (CI), as in (5).

$$P_i^P = \frac{\underline{\underline{a_i^P}}.\underline{W^P}}{W_i^P} \tag{4}$$

$$CI^P = \frac{\frac{\sum_{i=1}^{n} P_i^P}{n} - n}{n-1} \tag{5}$$

Next is to give a value to Saaty's random indexes (RI^P and RI^F). In the application of our ranking model, we worked with the interpolation of the Donegan-Dodd experimental results [21]. Based on the random indexes, the consistency ratios can be calculated (CR^P and CR^F), as in (6). If the CR values are bigger than the experimental 0.1 value, than the examined matrices are consistent.

$$CR^P = \frac{CI^P}{RI^P} \tag{6}$$

Next step of the AHP-based ranking is to calculate the preference ratios from the actual values. For this, we need to collect the actual values ($T_i^{P,z}$ and $T_k^{F,z}$) of the given criteria for all examined zones (Z_1, Z_2, … Z_w, in case of „w" examined zones), where „z" is an integer index ($z = 1 \ldots w$), which belongs to the examined zone. Then, the preference rations ($P_{z,i}^P$ and $P_{z,f}^K$) can be calculated with this formula (in the default case, the smaller actual value is the bigger; if the bigger value is the better one, we need to divide the actual value by the maximum of the actual values), as in (7), where $T_{i,MIN}^P = MIN\left\{T_i^{P,1}; T_i^{P,2}; \ldots; T_i^{P,w}\right\}$. Last step of the ranking is to calculate the ranking values for all examined zones (P_z^P and P_z^F), as in (8).

$$P_{z,i}^P = W_i^P \cdot \frac{T_{i,MIN}^P}{T_i^{P,z}} \tag{7}$$

$$P_z^P = \sum_{i=1}^{n} P_{z,i}^P [\%] \tag{8}$$

These ranking values will give us the current city logistics state $\left(P_z^P\right)$ and the future development potential $\left(P_z^F\right)$ of every examined zone. The bigger the calculated value is, the better is the current state or the development potential of the given zone, and the maximum of the values is 100%.

4.2 Evaluation Criteria

On Fig. 4, our purposes, the groups of the criteria, the criteria and the examined zones can be seen. We defined four main categories for the criteria (based on their sources).

First, the criteria were defined by us, but after the evaluation of them performed by experts, we added some new criteria based on the opinions of the experts [21]. In case of the examination of the current city logistics state, we worked with 43 criteria. 12 of them belong to the "questionnaire data" group, e.g. *the compliance of the actual entry regulations, the aspects of the logistics units, delivery times, cargo bike deliveries, outbound deliveries, warehouses, ERP system and home deliveries*. 12 other criteria belong to the "simulation-based data" group, e.g. *the number of transactions, the mileage, the performance, the emission, the stocks and the operation costs*. The 19 remaining criteria belong to the "network and regulation properties"

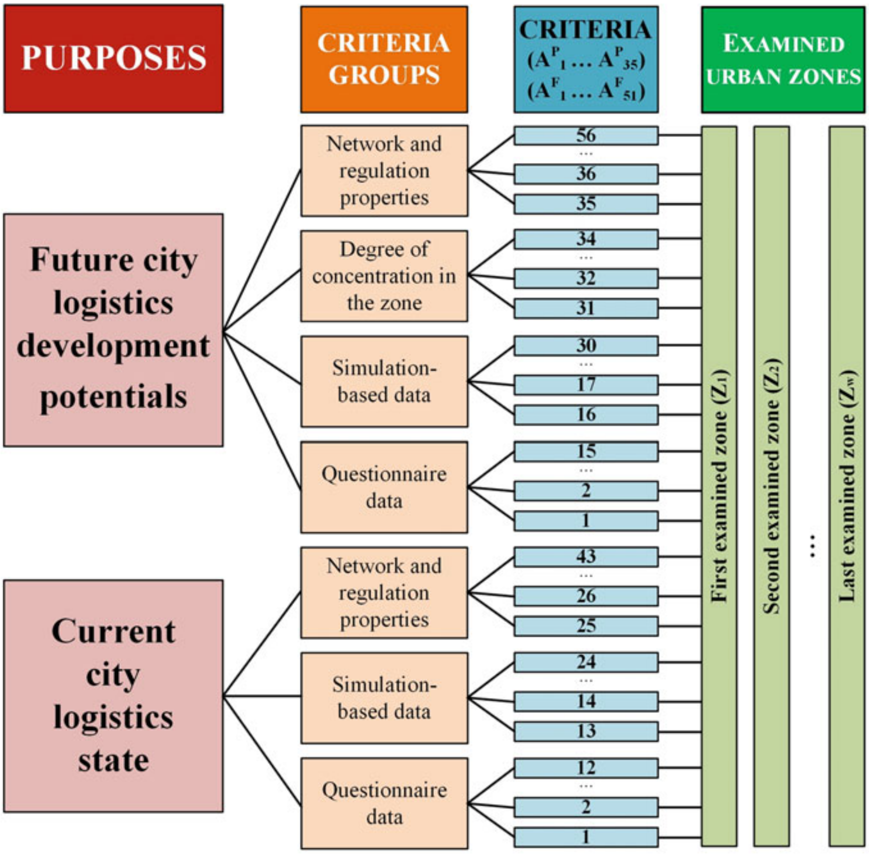


Fig. 4 Decision tree of the AHP-based ranking model

group, e.g. *the aspects of entry regulations, the loading possibilities, the state of the infrastructure, the aspects of electric charging, pick points and pedestrian couriers, and the possibilities and share of cargo bike, urban railway, waterway, combined, autonomous or drone deliveries.*

In case of the examination of the future development potentials, we worked with 56 criteria. 15 of them belong to the "questionnaire data" group, e.g. *the aspects of the logistics units, the groups of the delivered goods, delivery times, share of special goods, cargo bike deliveries, outbound deliveries, warehouses and ERP system, and the willingness of the delivery locations to participate in new city logistics systems.* 15 other criteria belong to the "simulation-based data" group, e.g. *the reduction or change of the number of transactions, the mileage, the performance and the emission, the aspects of the stocks, the operation costs and the investment costs.* 4 criteria belong to the "degree of concentration" group: *the single degree of concentration in delivery locations/km^2 area, the area proportional degree of concentration, in m^2 floor area of the delivery locations/m^2 floor area, and their special versions with considering the multilevel CSDLs.* The 22 remaining criteria belong to the "network and regulation properties" group, e.g. *the aspects of entry regulations, the loading possibilities, the state of the infrastructure, the aspects of electric charging, pick points and pedestrian couriers, the possibilities of cargo bike, urban railway, waterway, combined, autonomous or drone deliveries, the available brownfield lands for logistics purposes, the average distance from them and the necessary number of transshipment points in the new system.*

In the next step, we weighted these criteria with experts from the field of city logistics and urban transportation.

4.3 Criteria Evaluation by Experts

For the weighting of the criteria, we asked 18 experts, from the Faculty of Transportation Engineering and Vehicle Engineering the Budapest University of Technology and Economics (especially from the City Logistics Research Group [22]), from the Clean Air Action Group [23] and from the Centre for Budapest Transport (BKK) [24]. In the first round of the evaluation, we received 13 answers, we used these weights in the first tests (where we worked with 35 and 51 criteria). In this first round, we received some advices from the experts to add new criteria (8 for the current state and 5 for the development potentials), so we asked these 18 experts again to weight the new criteria, and we received 9 answers. Based on these answers, we calculated the final weights based on a weighting method. In the evaluation phase, the experts could choose the 4 categories for all criteria (see on Fig. 5).

At the weighting, we considered the experience of the experts in the field of city logistics and urban transportation. Based on the answers and the "experience categories", we calculated ranking values for the criteria (with considering the maximum sum value of the weights). Based on these values, we ranked the criteria, and based on the shares of the answers, we decided about the final weights. The occurrence of

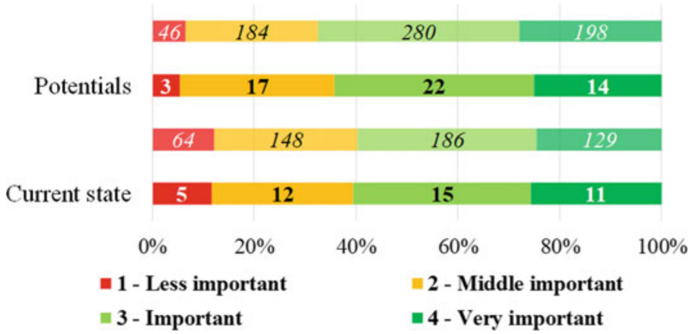


Fig. 5 Occurrence of the categories

the final categories can be seen on Fig. 5 with bold letters (data with italic letters shows us the occurrence of the answers of the experts).

5 Testing the Model for Shopping Malls in Budapest

As we had data and a mesoscopic simulation model for four shopping malls from Budapest, we decided to test our ranking model on these CSDLs, so in the testing phase, these four malls defined the four examined zones for us. The full process of the application of our AHP-based multicriterial ranking model can be seen on Fig. 6. In the testing, the first step was to collect the necessary data for the ranking: we performed analysis on database of our former data collection, we customized our simulation model and performed the simulation runs after experiment design, and we collected all other necessary data from other sources (e.g. research documents, bike path network map or electric charging network databases). These data were added to our ranking model as input data and we ranked the examined zones (shopping malls) by use of our ranking model.

5.1 Test Without a Fictive Zone

First, we performed the tests only for the examined shopping malls. Based on the results of these tests, the examined zones have a middle or high current city logistics state (with values between 60 and 77%), and all of them have high development potentials (with values between 69 and 83%). The main problem with these results, that our data collection [16] showed us earlier, that none of the examined malls is really well developed (in the viewpoint of city logistics), we discovered several development potentials, and none of the actual regulations and the actual network is well developed. In the case of the AHP models, we compare the zones only with

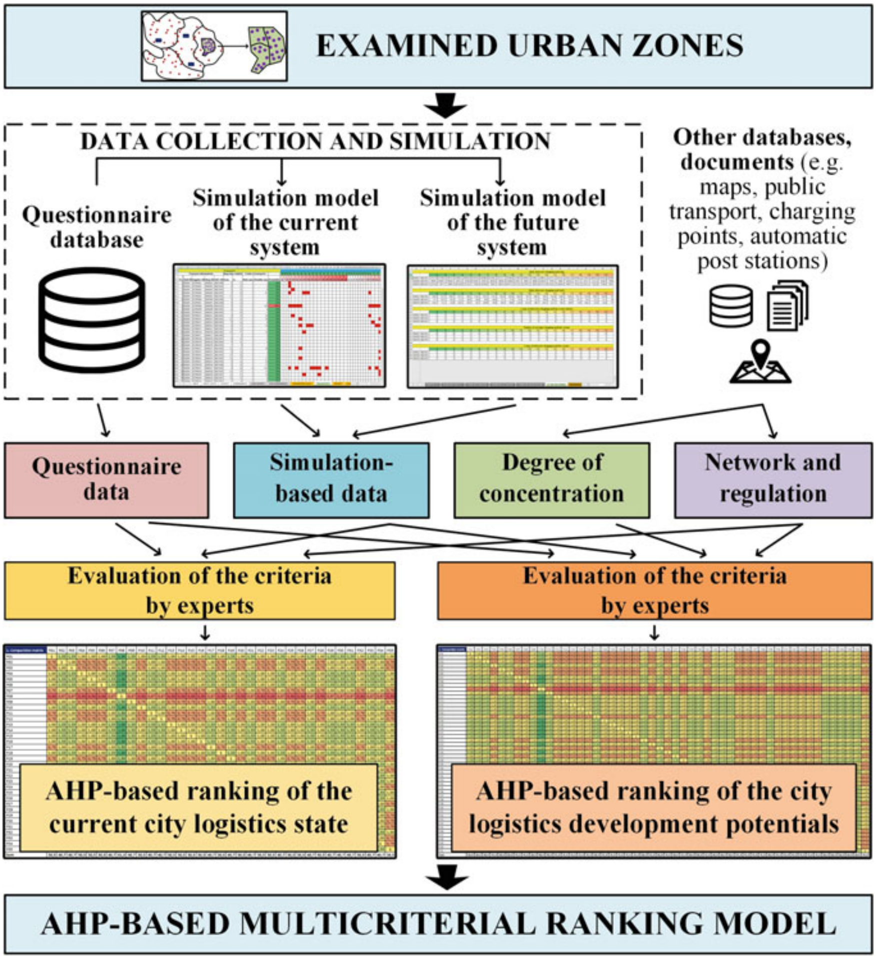

EXAMINED URBAN ZONES
DATA COLLECTION AND SIMULATION
Questionnaire database
Simulation model of the current system
Simulation model of the future system
Other databases, documents (e.g. maps, public transport, charging points, automatic post stations)
Questionnaire data
Simulation-based data
Degree of concentration
Network and regulation
Evaluation of the criteria by experts
Evaluation of the criteria by experts
AHP-based ranking of the current city logistics state
AHP-based ranking of the city logistics development potentials
AHP-BASED MULTICRITERIAL RANKING MODEL

Fig. 6 Application process of the AHP-based ranking model

each other, so here, we can compare our zones only to the best one of the middle-developed zones, so our results are distorted upwards. The ranking sequence of the zones will be correct in this case as well, but the final values won't be real. Based on this, we decided to add a so-called fictive zone for the next test.

5.2 Test with a Fictive Zone

We added the fictive zone (in case of these tests basically a fictive shopping mall) to our ranking model, to make it possible to compare the examined zones with a theoretically well-developed zone, with high development potentials. At the fictive zone, in case of values on a scale, we chose the best (the smallest or the biggest) value; in case of percentage values with exact optimum, we chose the best (0% or 100%) value; and in case of other values without an exact optimum, we chose the value, which is 10% better (smaller or bigger) than the best one of the examined zones. After adding the fictive zone, we performed the AHP-based ranking for five zones. The results can be seen on Fig. 7 and in Table 2 (with the same colors for the zones).

These results are a better reflection of reality. The fictive zone was the best one in case of every criteria, and the four examined zones are middle developed, with middle development potentials, so we received the expected results. Based on these results, we can say that it will be better to use the solution with a fictive zone in the future, as it gives as more realistic results.

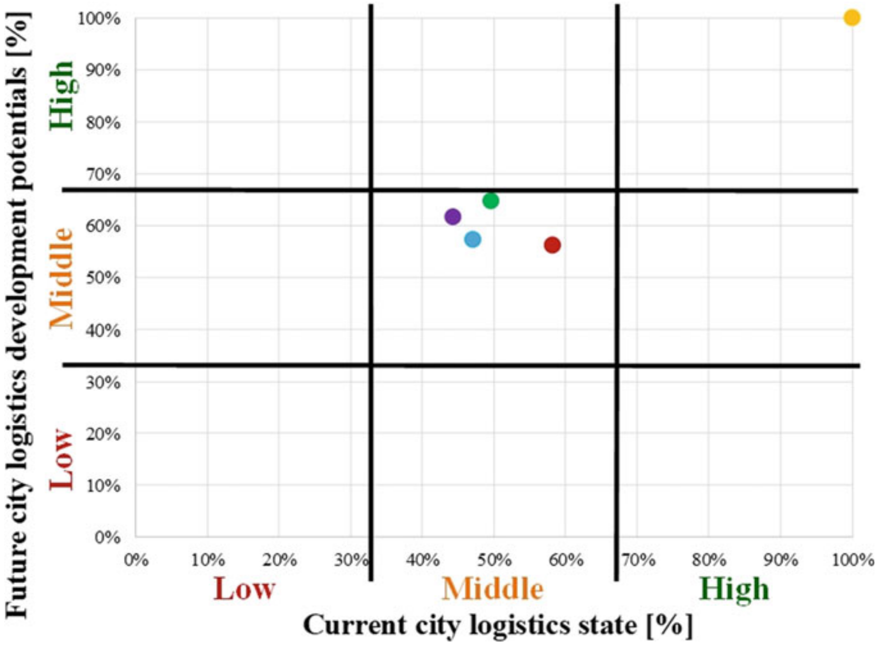


Fig. 7 Results of the ranking test

Table 2 Results of the ranking for four shopping malls

	P_z^P	P_z^F
SM001	49,62%	64,71%
SM002	47,12%	57,22%
SM003	58,26%	56,11%
SM004	44,36%	61,53%

6 Next Steps of the Research

We already defined several tasks to continue our research. The most important task is to test our model on the zones of the ‘Váci utca’ shopping area (in Budapest) as well, because we have data about its delivery locations, we examined the stores by our questionnaire and we developed a topological model for the whole area [17]. If the simulation model of the area will be ready, we are going to be able to rank 4 shopping malls and the zones of a shopping area with our AHP-based model. The analysis of a market from Budapest is also ongoing, with this, it will be possible to rank 9 zones by use of our model.

Naturally, as the investigation of the concentrated sets of delivery locations is in the focus of our research group, we have data and simulation models about this kind of special delivery locations, this made it possible to test our ranking model. In the future, we would like to add normal urban zones with single delivery locations to our model and test it with them too. It will be also a very important task in the future to perform some sensitivity analysis to examine the effects of changing the weights in our AHP-based ranking model.

7 Summary

In our paper, our main purpose was to present an AHP-based, two-dimensional multicriterial ranking model, which makes it possible to evaluate the current city logistics state and the future city logistics development potentials of urban zones. We expect, that this ranking model will be able to help future city logistics projects with highlighting those urban zones, they can be developed well. The literature review showed us, that our two-dimensional purposes are completely new, there was no ranking model developed for these purposes in city logistics. We decided to use the AHP method, we presented its formalization and the criteria. For the weighting, we asked 18 experts from the field of city logistics and urban transportation. Based on their answers and based on the degree of experience of the experts, we calculated the final weights of the criteria. By use of these final weights and by use of all other input data, we performed the tests of the multicriterial model for 4 urban zones (4 shopping malls). In the first tests, we examined only these zones, but the results were not appropriate, so we added a so-called fictive zone, and this helped us to get realistic results, the ranking model showed us correctly the current state and the development potentials too. At the end of our paper, we presented some tasks for the next steps of our research, but now we can clearly say, that we developed such a multicriterial ranking model, which will be able to rank urban zones in case of future city logistics projects.

Acknowledgements Supported by the ÚNKP-19-3 New National Excellence Program of the Ministry for Innovation and Technology.

References

1. K. Bóna, D.L. Sárdi, Analysis and mesoscopic modelling of logistics systems of concentrated urban delivery points—Koncentrált városi igénypontok áruellátó logisztikai rendszereinek elemzése és mezoszkópikus szintű modellezése. Logisztikai Évkönyv **2019**, 121–130 (2018)
2. K. Bóna, D. L. Sárdi, New concepts of logistics systems of the urban concentrated sets of delivery points, by use of different transportation modes—A városi koncentrált igénypont-halmazok áruellátási rendszerének új koncepciói a különböző közlekedési alágazatok lehetőségeinek kihasználásával *in XIII. IFFK Conference (Innovation and Sustainable Surface Transport)*, Budapest (2019)
3. J. Antún, R. Alarcón, Ranking Projects of logistics platforms: A methodology based on the electre multicriteria approach. Procedia—Soc. Behav. Sci. **160**, 5–14 (2014)
4. D. Patier, M. Browne, A methodology for the evaluation of urban logistics innovations. Procedia—soc. Behav. Sci. **2**(3), 6229–6241 (2010)
5. T. Teoh, O. Kunze, C. Teo, Methodology to evaluate the operational suitability of electromobility systems for urban logistics operations. Transport. Res. Proced. **12**, 288–300 (2016)
6. A. Awasthi, S. Chauhan, A hybrid approach integrating affinity diagram, AHP and fuzzy TOPSIS for sustainable city logistics planning. Appl. Math. Model. **36**(2), 573–584 (2011)
7. S. Tadić, S. Zeevi, M. Krstić, Sustainability of the city logistics initiatives, *in 3rd Logistics International Conference (LOGIC 2017)*, Belgrade (2017)
8. S. Tadic, S. Zecevic, M. Krstic, A novel hybrid MCDM model based on fuzzy DEMATEL, fuzzy ANP and fuzzy VIKOR for city logistics concept selection. Expert Syst. Appl. **41**, 8112–8128 (2014)
9. M. Gogas, G. Adamos. E. Nathanail, Assessing the performance of intermodal city logistics terminals in Thessaloniki. Transport, Res. Proced. **24**, 17–24, 2017
10. Y. Kara, Measuring the sustainability of cities in turkey with the analytic hierarchy process. Open J. Soc. Sci. **7**, 322–334 (2019)
11. M. Janjevic, D. Knoppen, M. Winkenbach, Integrated decision-making framework for urban freight logistics policy-making. Transport. Res. Part D: Transport Environ. **72**, 333–357 (2019)
12. J. Wątróbski, K. Małecki, K. Kijewska, S. Iwan, A. Karczmarczyk, R. Thompson, Multi-criteria analysis of electric vans for city logistics. Sustainability **9**(8), 1–34 (2017)
13. L. Zhao, H. Li, M. Li, Y. Sun, Q. Hu, S. Mao, J. Li, J. Xue, Location selection of intra-city distribution hubs in the metro-integrated logistics system. Tunn. Undergr. Space Technol. **80**, 246–256 (2018)
14. A. Awasthi, S. Chauhan, S. Goyal, A multi-criteria decision making approach for location planning for urban distribution centers under uncertainty. Math. Comput. Model. **53**(1–2), 98–109 (2010)
15. R. Saaty, The analytic hierarchy process—what it is and how it is used. Mathe. Modell. **9**(3–5), 161–176 (1987)
16. B. Mészáros, D.L. Sárdi, K. Bóna, Monitoring, measurement and statistical analysis (MMSA) based methodology for improvement city logistics of shopping malls in Budapest. World Rev. Int. Transport. Res. **6**(4), 352–371 (2017)
17. D.L. Sárdi, K. Bóna, Examination of the logistics systems of concentrated sets of urban delivery points by simulation, *in The 21th International Conference on Harbor, Maritime & Multimodal Logistics Modelling and Simulation*, Lisbon, 2019, pp. 1–10
18. D.L. Sárdi, K. Bóna, Developing a mesoscopic simulation model for the examination of shopping mall freight traffic in Budapest, *in Smart Cities Symposium 2017,* Prague (2017)
19. K. Bóna, Á. Róka, D.L. Sárdi, Mathematical modelling of the cost structure of the logistics system of shopping malls in budapest. Period. Polytechn. Transport. Eng. **46**(3), 142–150 (2018)

20. D.L. Sárdi, K. Bóna, Macroscopic simulation model of a multi-stage, dynamic cargo bike-based logistics system in the supply of shopping malls, in *Budapestin Smart Cities Symposium 2018* (Prague, 2018)
21. H. Donegan, F. Dodd, A note on saaty's random indexes. Math. Comput. Model. **15**(10), 135–137 (1991)
22. City Logistics Research Group of the Department of Material. *Budapest University of Technology and Economics*, 2020. Available: https://www.logisztika.bme.hu/citylog/. Accessed: 18 March 2020
23. Clean Air Action Group, *Levegő Munkacsoport* (2020). Available: https://www.levego.hu/en/. [Accessed: 2020. 03. 27.]
24. Centre for Budapest Transport (BKK). *Budapesti Közlekedési Központ*, 2020. Available: https://bkk.hu/en/news/. Accessed: 27 March 2020

The Influence of Religious Tradition and Social Trust on Corporate Cash Holdings

Zhijia Peng, Ming Xiao, and Lin Mu

Abstract We examine the impact of religious traditions and social trust on corporate cash holdings from 2007 to 2015 with the listed companies in China as the research sample. The empirical results show that in the areas where the traditional religion and social trust are higher, the cash holding of listed companies is lower, and social trust has the interaction enhancing effect on the impact of religious traditions on corporate cash holdings. Compared with the state-owned listed companies, the traditional religion and social trust have a negative impact on cash holdings in non-state-owned listed companies. Further research shows that the influence of traditional religion and social trust on the cash holdings of enterprises is mainly manifested as an over-curbing effect on the holding of cash. Traditional religion and social trust both as informal institutional factors, have complementary effect on formal institutional factors, such as the marketization process, the level of government governance and the level of legalization. A series of robustness tests prove the reliability of the above conclusion. This article provides empirical evidence that religious traditions and social trust influence the cash holdings of firms.

Keywords Cash holdings · Religious traditions · Social trust · Institutional environment · Nature of property rights

1 Introduction

Formal institutions such as government governance, law, and marketization have a vital role in promoting economic and social development. However, in different

Z. Peng (✉)
Aviation Industry Corporation of China, Ltd., Beijing, China
e-mail: pzj_pzj@126.com

M. Xiao · L. Mu
University of Science and Technology Beijing, Beijing, China
e-mail: xiaoming@ustb.edu.cn

L. Mu
e-mail: mulin704@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_5

regions where the level of the rule of law is relatively uniform, there are still major differences in various aspects such as social economy. Different regions have long-term historical development. The influence of informal systems such as religious traditions and cultural conventions accumulated cannot be ignored [1].

The influence of informal institutions on economic development has become a consensus in the academic community [2]. Religious traditions and social trust are important components of the informal system. The role played in the capital market gradually arises scholars' attention. Previous studies have shown that religious traditions can improve corporate governance, restrict corporate risk-taking behaviors, and reduce corporate risk-taking levels [3–5]; Social trust promotes stable social development and economic growth, and increases the company's production efficiency and market competitiveness [6, 7].

The issue of corporate cash holdings is one of the top ten challenges for the financial community. At present, the research on the influence factors of cash holdings is mostly focused on the characteristics of the micro-company and the meso-industry [8, 9]. There are also literatures that consider the economy environmental and formal institutional [10], and few literatures focus on the impact of informal systems such as religious traditions and social trust on corporate cash holdings.

Based on this, this article uses the financial data of Shanghai-Shenzhen A-share listed companies from 2007 to 2015 and provincial-level religious and social trust data to explore the impact of religious traditions and social trust on corporate cash holdings, and further consider different in the institutional environment.

The possible contributions of this paper are as follows: First, in our paper, the sociocultural and business ethics factors such as religious tradition and social trust are included in the discussion of the factors affecting the corporate cash holdings, which enriches the literature on the area of corporate cash holdings, and provides new ideas for related research. Second, from the micro level, we analyze the function mechanism of religious traditions and social trust affecting corporate behaviors, and provide empirical evidence for the impact of informal systems at the social level on micro-firm decisions. Third, based on the institutional background of China's transition economy, empirically examine the impact of formal institutional factors such as marketization process, government governance, and legalization on the socio-cultural role of religious traditions, social trust, etc., verifying that formal and informal institutions have a certain complementary influence on enterprises.

The following are arranged as follows: The second part reviews related literature and proposes research hypotheses; the third part is model design and variable definition; the fourth part is analysis of empirical results; the fifth part proposes research conclusions and recommendations of this paper.

2 Theoretical Analysis and Research Hypothesis

2.1 *Religious Traditions and Corporate Cash Holdings*

For a long time, studies on religion have focused on philosophical and sociological categories [11]. Since the 1990s, Western scholars have begun to pay attention to the influence of religion on corporate governance. Religion's influence on economic behavior mainly comes through two approaches: one is the shaping of personal value systems such as moral sentiment and identity of economic individuals, and the other is the restriction of external social norms and business ethics. With respect to corporate governance, studies have found that companies with strong religious atmosphere show stronger risk aversion and lower investment rates [3]. Corporate management is even more disgusted with litigation risk [4], relatively few financial reporting violations [5]. In recent years, some Chinese studies have also come to similar conclusions [1, 2, 12]. At present, there are few literatures to discuss the influence of religious traditions on corporate cash holdings. Based on analysis of other studies, religion's restraint on individual behavior of corporate management can reduce agency costs, and the regulation of external society can increase trust and reduce information asymmetry in the capital market [12], thus reducing the number of companies facing the financing constraints, in turn, affecting corporate cash holdings. Based on this, put forward hypothesis 1:

H1 The stronger the religious tradition of the place where the listed company is located, the lower the cash holdings of listed companies is.

2.2 *Social Trust and Corporate Cash Holdings*

In addition to material capital and human capital, "trust" is considered to be the main social capital that determines economic growth and social progress [13]. Adam Smith pointed out in his important book "The Theory of Moral Sentiments" that economic activities are based on social habits and morals. In the book "Trust: Creation of Social Morality and Prosperity," On the one hand, social trust can improve the quality of governance formal system, promote economic development and increase efficiency [7]. On the other hand, social trust can help to alleviate corporate agency problems and information asymmetry problems and improve financial markets [8], creating a fair and transparent trading environment, reducing the opportunity cost and moral hazard in the implementation of contracts [14], so as to jointly provide investors with stable psychological expectations [15], reducing the external cost of financing, in turn, weakens the company's preventive motives for holding cash. The dual institutional environment of emerging and transition in China has exacerbated the differences in regional levels of social trust [16]. Based on the above analysis, the second hypothesis of this paper is proposed:

H2 The stronger the social trust of the place where the listed company is located, the lower the cash holdings of the listed company is.

2.3 Religious Tradition, Social Trust and Corporate Cash Holdings

As an important part of the informal system, religious traditions and social trust have a close relationship with a society's business spirit and social culture. In the study of religious influence on economic activity, some analysts believe that when the behavioral norms of regional religious traditions are recognized by local social conventions, they will increase the level of mutual trust among random individuals in society [12]. Therefore, these two informal systems (religious traditions and social trust) should be further combined to examine whether there is an interaction between their influence on corporate cash holdings. Because the religious tradition and social trust affect the mechanism of corporate cash holdings, there is a certain commonality. Both can alleviate the agency problems and financing constraints, and then weaken the preventive motives for holding cash. Therefore, it can be expected that religious traditions have a greater impact on corporate cash holdings in regions with higher social trust; In the same way, the influence of social trust is even more pronounced in regions with higher religious traditions. Therefore, hypothesis 3 is presented as follows:

H3 Religious tradition and social trust have a positive interaction effect on corporate cash holdings.

Combined with China's special institutional background, we further examine the relationship between the impact of informal institutions such as religious traditions and social trust on corporate cash holdings and the nature of corporate property rights. Compared with non-state-owned enterprises, state-owned enterprises are generally large-scale and have relatively small operating risks. Coupled with the attributes of their state-owned ownership, they often have easier access to external financing and face a relatively small degree of financing constraints [16]. Correspondingly, non-state-owned enterprises are faced with greater financing constraints and have greater development uncertainty and are more likely to hold excessive cash for preventive motives. Therefore, a higher level of informal institutions can significantly reduce the external financing costs of non-state-owned enterprises. It can be expected that the cash holdings of non-state-owned enterprises are more sensitive to religious traditions and social trust. Therefore, put forward the fourth hypothesis of this paper:

H4 Compared with state-owned listed companies, religious traditions and social trust have a greater impact on the cash holdings of non-state-owned listed companies.

Social institutions include formal and informal institutions [4] is concerned with countries with a relatively complete formal institutions, while China has a unified legal institution, but due to its vast territory and large interregional differences, the process of marketization, governance, and legalization among regions is inconsistent, there are major differences in the institution environmental quality in the provincial level [13]. In areas with high institutional environment quality, the internal and external information exchange mechanisms of enterprises are more developed, and they have a sound institutional guarantee and business environment, which may lead to the crowding out of informal institutions such as religious beliefs and social trust. The role is limited [15]. When the formal institution provides weaker protection for investors, it is more likely to have a moral crisis. Investors whether trust management or not has a direct influence on the formulation of their investment decisions [16]. Therefore, an increase in the level of informal institutions may increase investors' willingness to invest, thereby reducing the financing constraints companies face, and thus reducing the company's behavior of holding excess cash out of preventive motives. Based on the above analysis, it can be expected that informal institutions such as religious traditions and social trust may play a more important role in a market where the formal system is less protective of investors. Based on this, hypothesis 5 is put forward:

H5 The weaker the formal protection of investors in the market, the more significant the impact of religious traditions and social trust on corporate cash holdings.

3 Study Design

3.1 Sample Selection

We select the listed companies on the China A-Shares Main Board from 2007 to 2015 as the original research samples. The company's data are from the CSMAR databases and Wind databases. Since China adopted a new version of the cash flow statement in 2007, in order to avoid inconsistency in the data caliber, the starting year for this study sample is 2007. We adopt the "Guidelines for Industry Classification of Listed Companies" issued by the China Securities Regulatory Commission in 2012 to divide the industry. According to the research needs, the sample companies are selected as follows: Excluding financial listed companies; Excluding listed companies with abnormal financial indicators, including ST and PT listed companies with negative operating income and continuous losses; Exclude companies that issued shares in A, B or H shares at the same time; Remove the company with missing major variables. In order to reduce the impact of outliers, STATA 14.2 is used to shrink the main continuous variable of listed companies by 1%.

3.2 Definition of Variables

1. *Explained Variable*

   (a) *Listed company cash holdings (cash)*

For the measurement of cash holdings, there are three main ways to define the cash holdings: Cash and cash equivalents/total assets [17]; Cash and cash equivalents/net assets [9], among them, net assets = total assets —ash and cash equivalents. The cash holding ratio calculated in this way will appear to be greater than 1 and often requires the use of the first way for robustness testing; Cash and cash equivalents/operating income [18]. This article uses the first way that is widely used in research.

2. *Explanatory Variables*

   (a) *Religious tradition (religionZX)*

This article refers to [1], we manually collect the proportion of members of religious circles in the Chinese People's Political Consultative Conference (CPPCC) of each province as the proxy variable of religious tradition. The source of the data is the list of the members of the CPPCC provincial committees from 2007 to 2016. According to China's political consultation system, the CPPCC members are composed of representatives from industries and fields in which people from all walks of life are located. Therefore, the higher the proportion of religious figures in the CPPCC National Committee members, the more important the religious activities in the province [1]. At the same time, using the documents approved by the State Council of China in 1983: "The report of the State Council Religious Affairs Bureau on determining the concept of Buddhism temples and Taoism temples in the Han area", the annex lists the list of key Buddhist temples in China's Han area and has 148 provincial key temples. According to the religious venues awarded by the National Religious Affairs Bureau in 2010 for the "First National Building of Advanced Collectives and Advanced Individuals of the Temple of Harmony Temples," we use this data to measure religious traditions in the robustness test.

(b) *Social trust (trustCGSS)*

This article refers to [19], using the survey data measured by the China General Social Survey (CGSS) in 2003 and 2010 as a proxy variable for social trust. According to the questionnaire, the average level of trust of residents in the provinces to strangers was calculated. The greater the value, the higher the level of trust in strangers. At the same time [14] entrusted the "China Entrepreneur Survey System" to a nationwide questionnaire survey in 2000 to draw the confidence indicators of each province, we used this indicator for robustness testing.

3. *Control Variables*

In order to control the influence of company characteristics on cash holdings, refer to the relevant literatures and empirically select and obtain the control variables of

Table 1 Main research variables and definitions

Variables	Definition and calculation method of variables
cash	Cash and cash equivalents/total assets
religionZX	The proportion of religious members in the provincial according to CPPCC
trustCGSS	CGSS survey data on social trust
tl	Total liabilities/total assets
capExp	Cash paid for acquisition of fixed assets, intangible assets and other long-term assets/total assets
cflow	Net cash flow generated from operating activities/total assets
tobin	Company's total market capitalization/total share capital
NetWC	(Current Assets—Current Liabilities—Currency Funds)/total assets
payout	Cash dividend per share * A shares outstanding/total assets
tshr	Circulating Shares/total shares

this paper. Including: asset ratio (tl), capital expenditure (capExp), cash flow (cflow), growth opportunity (tobin), net working capital (NetWC), cash dividend (payout), equity structure (tshr).

The definition methods for the main research variables in this paper are listed in Table 1.

3.3 *Model Setting*

(1) *1.Religious Traditions and Corporate Cash Holdings Model*

In this paper, based on [9] model of studying cash influencing factors, combining China's relevant research to select control variables that are in line with China's national conditions, and based on this, adding the explanatory variables of religion tradition (religionZX) to establish a model (1). Explore the impact of regional religious traditions on corporate cash holdings.

$$\begin{aligned} Cash_{it} = {} & \beta_0 + \beta_1 religionZX_{it} + \beta_2 tl_{it} \\ & + \beta_3 capExp_{it} + \beta_4 cflow_{it} + \beta_5 tobin_{it} \\ & + \beta_6 NetWC_{it} + \beta_7 payout_{it} + \beta_8 tshr_{it} + \varepsilon_{it} \end{aligned} \tag{1}$$

2. *Social Trust and Corporate Cash Holdings*

In order to explore whether social trust has an impact on corporate cash holdings, we replace the explanatory variables in model (1) with trustCGSS and establishes a model (2).

$$Cash_{it} = \beta_0 + \beta_1 trustCGSS_{it} + \beta_2 tl_{it}$$

$$+ \beta_3 capExp_{it} + \beta_4 cflow_{it} + \beta_5 tobin_{it} + \beta_6 NetWC_{it} + \beta_7 payout_{it} + \beta_8 tshr_{it} + \varepsilon_{it} \quad (2)$$

3. *Religious Tradition, Social Trust and Corporate Cash Holdings*

Further, we add the two informal institutional factors of religious tradition and social trust to the model at the same time, and adds the interaction terms between the two to establish the model (3) to examine whether they have an interaction effect on the impact of the company's cash holdings.

$$\begin{aligned} Cash_{it} = \beta_0 + \beta_1 religionZX_{it} + \beta_2 trustCGSS_{it} + \\ \beta_3 relig\_tru_{it} + \beta_4 tl_{it} + \beta_5 capExp_{it} + \beta_6 cflow_{it} + \\ \beta_7 tobin_{it} + \beta_8 NetWC_{it} + \beta_9 payout_{it} + \beta_{10} tshr_{it} + \varepsilon_{it} \end{aligned} \quad (3)$$

4 Empirical Results Analysis

4.1 Descriptive Statistics

Table 2 shows the descriptive statistics of the main research variables in this paper, with a total of 16,515 sample data. The average cash holdings (cash) of listed companies in the sample is 20.7%, the median is 15.9%, the minimum is 1%, the maximum is 73.9%, and the standard deviation is 0.158, indicating that the listed company's cash holdings is overall high, and there are large differences between companies.

Table 2 Descriptive statistics

Variable	N	Mean	P50	Sd	Min	Max
cash	16,515	0.207	0.159	0.158	0.0100	0.739
religionZX	16,515	0.0190	0.0160	0.0160	0.00,700	0.195
trustCGSS	16,515	3.056	3.043	0.144	2.836	3.590
market	16,515	7.398	7.560	1.780	−0.300	10.65
size	16,515	21.73	21.61	1.172	19.16	25.74
tl	16,515	0.441	0.439	0.222	0.0440	0.998
capExp	16,515	0.0570	0.0410	0.0540	0	0.256
cflow	16,515	0.0420	0.0410	0.0780	−0.200	0.255
tobin	16,515	2.840	2.167	2.116	0.918	13.45
NetWC	16,515	0.00700	0.0230	0.215	-0.631	0.490
payout	16,515	0.00800	0.00400	0.0120	0	0.0690
tshr	16,515	0.696	0.730	0.281	0.154	1

Table 3 Correlation test

Variables	Cash	Religi ~ X	TrustC ~ S	Tl	CapExp
cash	1				
religionZX	−0.056***	1			
trustCGSS	−0.112***	0.197***	1		
tl	−0.486***	0.065***	0.092***	1	
capExp	−0.046***	0.021***	0.031***	−0.112***	1
cflow	0.126***	0.016**	0.016**	−0.128***	0.156***
tobin	0.210***	0.00200	−0.024***	−0.265***	−0.035***
NetWC	0.119***	−0.047***	−0.091***	−0.570***	−0.106***
payout	0.130***	−0.031***	−0.031***	−0.292***	0.069***
tshr	−0.356***	0.035***	0.041***	0.316***	−0.168***

Variables	cflow	tobin	NetWC	payout	tshr
cflow	1				
tobin	0.097***	1			
NetWC	−0.191***	0.092***	1		
payout	0.245***	0.096***	0.169***	1	
tshr	0.029***	−0.151***	−0.191***	0.140***	1

The mean of religious traditions (religionZX) is 1.9%, which means that the average proportion of religious people in the CPPCC is 1.9%. The lowest proportion in the province is 0.7% (Sichuan) and the highest is 19.5% (Tibet). These can better represent the influence and status of religious traditions in each province. The mean of social trust (trustCGSS) is 3.056, the lowest is 2.836 (Tianjin), the highest is 3.590 (Ningxia), and the standard deviation is 0.144, which reflects the difference in social trust among the provinces. See Table 2 for the descriptive statistics of the remaining variables.

This article uses the Pearson correlation coefficient method to check the correlation between variables before regression. The correlation coefficient is shown in Table 3, in which corporate cash holdings (religion ZX) and social trust (trustCGSS) are both significantly negative. Relevant, preliminary explanation of the influence of religious traditions and social trust on corporate cash holdings is negative.

4.2 Analysis of Basic Regression Results

Table 4 shows the basic regression results of this paper, both using the standard error of robustness OLS method to estimate. The first column is the regression result of model (1), and the explanatory variable is the religious tradition; the second column is the regression result of model (2), the explanatory variable is social trust; the third

Table 4 Basic regression results

	(1)	(2)	(3)	(4)	(5)
Variables	Religion	Trust	Religandtrust	State-owned	Non-state-owned
religionZX	−0.288**		−11.110***	−6.276	−13.397***
	(−2.25)		(−3.65)	(−1.49)	(−4.03)
trustCGSS		−0.090***	−0.149***	−0.113***	−0.152***
		(−6.06)	(−6.20)	(−3.07)	(−5.31)
relig_tru			3.448***	1.960	4.195***
			(3.61)	(1.48)	(4.04)
tl	−0.493***	−0.492***	−0.492***	−0.257***	−0.582***
	(−33.48)	(−33.53)	(−33.51)	(−11.81)	(−31.09)
capExp	−0.316***	−0.316***	−0.315***	−0.192***	−0.409***
	(−14.04)	(−14.02)	(−13.98)	(−5.54)	(−14.88)
cflow	0.039***	0.039***	0.040***	0.128***	0.025
	(2.72)	(2.74)	(2.78)	(6.33)	(1.33)
tobin	−0.002**	−0.002**	−0.002**	0.003**	−0.003***
	(−2.23)	(−2.26)	(−2.30)	(1.97)	(−3.64)
NetWC	−0.315***	−0.316***	−0.316***	−0.187***	−0.392***
	(−23.93)	(−24.00)	(−24.01)	(−10.59)	(−22.60)
payout	−0.297***	−0.297***	−0.298***	0.172	−0.209*
	(−2.97)	(−2.97)	(−2.98)	(1.01)	(−1.76)
tshr	−0.120***	−0.120***	−0.120***	−0.028***	−0.169***
	(−23.25)	(−23.25)	(−23.22)	(−3.91)	(−24.99)
N	16,515	16,515	16,515	6290	10,225
r2_b	0.385	0.393	0.396	0.160	0.451

column is the regression result of model (3), are add the interaction of religious traditions and social trust based on the models (1) and model (2). At the same time, we also take into account the differences in the nature of corporate ownership, and classifies the sample companies into two groups, state-owned and non-state-owned, according to the nature of property rights. The results are shown in columns (4) and (5) respectively.

From column (1) in Table 4, it can be seen that the regression coefficient of the religious tradition is significantly negative (−0.288), which verifies the hypothesis 1 of this paper that the influence of religious tradition on corporate cash holdings is negative. The stronger the religious tradition of the place where the company is located, the lower the company's cash holdings.

Column (2) in Table 4 shows that the regression coefficient of social trust is significantly negative at the 1% level, which is consistent with expectations. Hypothesis 2 is verified that social trust has a negative effect on the cash holdings of enterprises.

The higher the social trust in the company's location, the lower the company's cash holdings.

The two informal institutional factors, religious tradition and social trust, are added to the model, and the engagements between the two are added to examine the interaction between the two. According to the results in column (3) of Table 4, the regression coefficients of religious traditions and social trust are both significantly negative at the 1% level, but the coefficients of the two are no longer important after the addition of the interline item. Our focus is on regression coefficients of interline item. The regression coefficient of the interaction between the two items (relig_tru) on corporate cash holdings is significantly positive at the 1% level (3.448), indicating that there is a positive interaction between religious tradition and social trust on the impact of cash holdings, that is, in regions with higher trust, religious traditions have a greater impact on corporate cash holdings. Similarly, in regions with higher religious traditions, the influence of social trust is even more pronounced. Hypothesis 3 is verified.

Combining the status of the socialist market economy with Chinese characteristics, state-owned enterprises and non-state-owned enterprises have great heterogeneity, further grouped regression analysis based on distinguishing the nature of property rights (column (4) and column (5)). We find that in the sample of state-owned listed companies, the coefficient of the religious item (relig_tru) is positive but no longer significant. In the sample of non-state listed companies, the coefficient of the intersection factor is still significantly positive at the 1% level, and the absolute value is greater than the sample group and the sample group of state-owned listed companies. This shows that compared with state-owned listed companies, religious traditions and social trust have more significant impact on the cash holdings of non-state-owned listed companies, which verifies the hypothesis 4.

4.3 Further Analysis

1. *Inhibition of Excessive Cash Holdings*

Cash holdings, as an important financial decision for a company, are related to the survival and development of the company. If the cash holdings are too low, there may be situations such as the breakage of the capital chain, or even financial difficulties. If the cash holdings are too high, the investment efficiency of the enterprise will be reduced, and the growth of the company will be restricted. Therefore, according to the trade-off theory, there is an optimal cash holding, so that the marginal cost of the company holding cash is equal to the marginal benefit.

In order to explore whether the influence of informal institutions such as religious traditions and social trust on corporate cash holdings are mainly reflected in the suppression of excessive cash holdings by firms. In this paper, the sample companies are grouped by year and industry to obtain the median annual cash holdings of various industries, and the difference between the actual cash holdings and the median value

(Dcash) of the company is calculated. Sample companies above the median cash holding size (Dcash > 0) are defined as high cash in cash, and sample companies below the median (Dcash < 0) are defined as undercash.

In model (1), interaction items (rel_highcash) of religious traditions and excessive cash holding is added to examine the inhibitory effects of religious traditions on firms holding excess cash. The regression results are shown in column (1) of Table 5. In model (2), interaction items (tru_highcash) of social trust and excessive cash holding is added to examine the effect of social trust on the holding of excess cash by enterprises. The results are shown in column (2) of Table 5. Next, the sample companies will be divided into cash over-holding and under-cash holdings, regression

Table 5 Inhibition of enterprises holding excessive cash

	(1)	(2)	(3)	(4)
Variables	Cash	Cash	Dcash > 0	Dcash < 0
religionZX	−2.262***		−11.881***	−1.719
	(−6.23)		(−4.27)	(−1.24)
trustCGSS		−0.066***	−0.135***	−0.034***
		(−9.75)	(−5.94)	(−2.88)
rel_highcash	25.294***			
	(5.13)			
tru_highcash		0.371***		
		(175.81)		
relig_tru			3.708***	0.510
			(4.27)	(1.18)
m	−0.319***	−0.089***	−0.552***	−0.049***
	(−11.17)	(−15.42)	(−34.56)	(−8.27)
capExp	−0.164***	0.062***	−0.608***	0.073***
	(−5.26)	(6.49)	(−21.42)	(6.29)
cflow	0.033**	0.010	−0.004	0.026***
	(2.52)	(1.61)	(−0.21)	(3.20)
tobin	−0.002***	−0.002***	−0.000	−0.001**
	(−3.12)	(−7.51)	(−0.46)	(−2.48)
NetWC	−0.199***	−0.034***	−0.426***	−0.012***
	(−10.13)	(−7.35)	(−25.58)	(−2.61)
payout	−0.068	0.161***	−0.463***	0.162***
	(−0.72)	(4.14)	(−3.81)	(2.75)
tshr	−0.084***	−0.019***	−0.131***	−0.011***
	(−9.33)	(−8.66)	(−21.03)	(−3.86)
N	16,515	16,515	8227	8227
r2_b	0.614	0.871	0.416	0.025

of model (3) to further explore whether religious traditions and social trust impact on corporate cash holdings is mainly reflected in the suppression of excess cash. The results are shown in column (3) and column (4) of Table 5.

In Table 5, the regression results in column (1) indicate that, the regression coefficient of interaction items (rel_highcash) is significantly positive at the 1% level, indicating that the higher the company's cash holding, the more religious traditions are significant inhibition of cash holdings. Column (2) shows that the regression coefficient of the interaction term (tru_highcash) is also significantly positive at the 1% level, indicating that the higher the company's cash holdings, the greater the inhibition of corporate trust on corporate cash holdings. The regression coefficient for the interaction term (relig_tru) in column (3) (excess cash holding group) is significantly positive at the 1% level; the regression coefficient of the interaction term (relig_tru) in column (4) (the cash holding group) is not significant; this shows that the impact of religious traditions and social trust on corporate cash holdings is mainly reflected in the inhibition of excess cash. The empirical results support the trade-off theory that firms have a target cash holding.

2. *The Complementary Functions of Informal Institutions and Formal Institutions*

According to Wang Xiaolu and Fan Gang's "China's Marketization Index by Provinces (CMI) (2016)" [20], the marketization index, the scores of relations between the government and the market, and the scores of development of market intermediary organizations and legal institution environmental, we use these three indexes as agency variables of marketization level (Mkt), government governance level (Gov), and legalization level (Law) respectively. CMI includes data from 31 provinces from 2008 to 2014, and uses 2008 as the base data. In our paper, we use the data of the last two periods to fit the index values of 2007 and 2015 respectively. According to the measurement method given by CMI, the larger the values of Mkt, Gov, and Law, the higher the level of regional marketization, government governance, and legalization. Therefore, we calculate the median of the three variables Mkt, Gov, and Law according to industry and annual groupings. According to this, the sample companies were divided into two groups based on the level of the three formal systems, a total of six groups, and the regression results for the model (3) are shown in Table 6.

From the columns (1) and (2) of Table 6, it can be seen that the interaction item (relig_tru) is not significant in the higher market level group, but is significantly positive in the lower marketization level group. This shows that in areas where the marketization process is slow, informal institutional factors such as religious traditions and social trust have a more significant impact on corporate cash holdings. The same result also appears in the regression of the grouping by the level of government governance and the rule of law. The empirical results show that there is a complementary relationship between formal and informal institutions' impact on corporate cash holdings, which validates hypothesis 5 of this paper. That is, the weaker the formal protection of investors in the market, the more obvious the impact of religious traditions and social trust on corporate cash holdings.

Table 6 Complementary roles of informal institutions and formal institutions

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Mkt_h	Mkt_l	Gov_h	Gov_l	Law_h	Law_l
religionZX	−25.814	−7.280**	−7.030	−9.058***	−13.016	−8.022***
	(−1.42)	(−2.45)	(−0.53)	(−2.94)	(−0.68)	(−2.63)
trustCGSS	−0.194**	−0.121***	−0.113	−0.132***	−0.153	−0.138***
	(−2.01)	(−4.65)	(−1.58)	(−4.88)	(−1.63)	(−5.13)
m	7.421	2.279**	1.676	2.820***	3.490	2.514***
	(1.25)	(2.44)	(0.39)	(2.92)	(0.57)	(2.63)
tl	−0.549***	−0.426***	−0.532***	−0.433***	−0.549***	−0.443***
	(−25.50)	(−24.26)	(−24.88)	(−24.69)	(−25.25)	(−24.15)
capExp	−0.427***	−0.268***	−0.421***	−0.274***	−0.411***	−0.263***
	(−12.80)	(−9.02)	(−12.62)	(−8.96)	(−12.11)	(−9.10)
cflow	0.047**	0.041**	0.037*	0.041**	0.039*	0.054***
	(2.31)	(2.09)	(1.76)	(2.16)	(1.93)	(2.84)
tobin	−0.003***	−0.000	−0.002*	−0.000	−0.003***	−0.001
	(−2.97)	(−0.41)	(−1.73)	(−0.41)	(−3.09)	(−0.87)
NetWC	−0.360***	−0.263***	−0.361***	−0.265***	−0.353***	−0.278***
	(−17.73)	(−16.92)	(−18.42)	(−17.65)	(−17.29)	(−17.43)
payout	−0.345**	−0.098	−0.363**	−0.157	−0.407***	−0.129
	(−2.38)	(−0.75)	(−2.45)	(−1.19)	(−2.71)	(−1.00)
tshr	−0.141***	−0.099***	−0.147***	−0.109***	−0.134***	−0.103***
	(−18.00)	(−14.90)	(−17.96)	(−16.00)	(−16.98)	(−15.33)
N	7369	9146	7543	8972	7048	9467
r2	0.395	0.355	0.368	0.360	0.419	0.371

4.4 Robustness Test

1. *Change the Regression Method*

Since the explained variable (cash holdings) in this paper is the ratio of the cash and equivalents to the total assets in the listed company, the value is distributed between 0 and 1, which is a restricted dependent variable. Therefore, in order to test whether the accuracy of the standard error OLS estimation method used in the paper is accurate, the three models of this paper are estimated using the panel tobit method. The estimation results are consistent with the results obtained by using the robustness standard error OLS method. The regression results of model (3) are listed in column (1) of Table 7.

Table 7 Robustness test regression results

	(1)	(2)	(3)	(4)	(5)
Variables	Tobit	Balance	Control	PSM in state-owner	PSM in non-state-owned
religionZX	−11.135***	−8.674**	−9.225***	−4.515	−13.479***
	(−3.41)	(−2.30)	(−3.19)	(−1.07)	(−3.17)
trustCGSS	−0.150***	−0.145***	−0.120***	−0.092**	−0.151***
	(−6.36)	(−4.61)	(−5.43)	(−2.54)	(−4.15)
relig_tru	3.453***	2.738**	2.912***	1.385	4.236***
	(3.37)	(2.30)	(3.21)	(1.04)	(3.20)
tl	−0.495***	−0.227***	−0.474***	−0.272***	−0.462***
	(−64.41)	(−13.95)	(−33.05)	(−11.65)	(−21.10)
capExp	−0.305***	−0.146***	−0.348***	−0.240***	−0.361***
	(−17.21)	(−5.32)	(−15.92)	(−5.71)	(−8.47)
cflow	0.039***	0.105***	0.079***	0.123***	0.088***
	(3.41)	(6.53)	(5.75)	(4.32)	(2.97)
tobin	−0.002***	0.004***	0.000	0.003*	−0.001
	(−4.24)	(3.62)	(0.51)	(1.85)	(−0.67)
NetWC	−0.321***	−0.193***	−0.320***	−0.177***	−0.259***
	(−44.41)	(−13.03)	(−24.63)	(−8.84)	(−13.51)
payout	−0.329***	0.536***	−0.119	0.094	−0.212
	(−4.02)	(4.34)	(−1.26)	(0.60)	(−1.06)
tshr	−0.119***	−0.025***	−0.129***	−0.043***	−0.121***
	(−33.76)	(−4.10)	(−23.26)	(−4.82)	(−13.96)
N	16,515	8730	16,515	3678	3590
r2		0.152	0.501	0.176	0.416

2. *Change the Sample Size*

In order to avoid sample selection biases, the research sample used in this paper is unbalanced panel data, including all A-share listed companies with full data from 2007 to 2015 after screening, with a total of 16,515 observations. Therefore, in order to test the robustness of the results, the sample size was changed here, and three models were regressed using the balance panel. There were 8730 observations for 9 years. The regression results for model (3) are listed in column (2) of Table 7. The coefficients of interaction terms of religious traditions and social trust is still significantly positive, but the significance level is reduced from 1% to 5%, it is still not changing analysis conclusion.

3. *Replace Proxy Variables For Explanatory Variables*

In terms of religious traditions, the “Provincial Key Temples” established by the State Council in 1983 and the “Rewards for Religious Places” selected by the National

Ethnic Religious Office in 2010 are used for robustness tests. In terms of social trust, we use the "China Entrepreneur Survey System" questionnaire survey conducted in 2000 to obtain a robustness test for each province's confidence index. It is still not changing analysis conclusion.

4. *Change the Control Variable*

Excluding the control variable "tobin" and "payout", which are not always significant in the above regressions, as well as further controlling the industry and the year based on the three models in this paper, the regression results are consistent with the above. The results of controlling industry and annual regression based on model (3) are listed in column (3) of Table 7.

5. *Matching Property Rights Using Propensity Score Matching (PSM) Method*

When discussing the differences in the influence of religious traditions and social trust on the cash holdings of state-owned listed companies and non-state-owned listed companies, the samples are simply divided into two groups according to the nature of the property rights, ignoring the differences between the two groups of samples, the endogenous problems caused by the sample selection bias are not resolved here. The PSM method is used to perform a 1:1 nuclear matching on the nature of property rights, and a sample with no significant difference between the two groups of control variables is obtained. The results of PSM regressions for model (3) are listed in columns (4) and (5) of Table 7 respectively, and are not significantly different from the above. Compared with state-owned listed companies, religious traditions and social trust have a more significant effect on the cash holdings of non-state-owned listed companies.

5 Conclusions and Recommendations

China has a wide geographical area. In the long-term development, different regions have great differences in religious traditions and social trust due to factors such as history and environment. We select listed companies on the A-share Main Board in China from 2007 to 2015 as research samples to examine the impact of religious traditions and social trust on corporate cash holdings. The study finds that: in regions where traditional religion and social trust are higher, the cash holdings of listed companies are lower, and the interaction between the two increases. Compared with state-owned listed companies, traditional religions and social trust have more significant impact on the cash holdings of non-state listed companies. Further research shows that the impact of traditional religion and social trust on corporate cash holdings is mainly manifested in the inhibition of excessive cash holdings. Traditional religion and social trust both as informal institutional factors, complement each other with formal institutional factors such as marketization processes, government governance levels, legalization levels, etc.

The sustainable development of the Chinese economy cannot be separated from the establishment of formal institutions and the cultivation of informal institutions. In this article, we combine religious traditions, social trust, and corporate cash holdings to explore and research, which is of important theoretical and practical significance to our understanding of the financial decisions of Chinese listed companies during the transition period. This has prompted us to think more fully and deeply about the role of informal institutions such as religious traditions and social trust in the development of business operations. For enterprises, it is necessary to constantly improve the internal institutions and establish a good communication mechanism with the government and the market. At the same time, they should dare to undertake social responsibilities and establish a good corporate image. For the government, it is necessary to improve the institutional environment, further deepen market-oriented reforms, create a fair and transparent economic environment, and promote social and economic development.

Due to data limitations, our study also has certain deficiencies. Academic research on religious traditions and social trust is often limited by the availability and authority of data. Although this article comprehensively considers various factors when selecting proxy variables and tries to avoid measuring errors, it is inevitably limited. It is expected that the data and methods in the future will become more and more perfect, and more detailed studies will be conducted on related fields.

References

1. D. Chen, X. Hu, S. Liang, F. Xin, Religious tradition and corporate governance. Econom. Res. J. **543**(9), 71–84 (2013)
2. H. Xinping, D. Yiyi, W. Ruoyu, Traditional religion, market liberalization and corporate risk-taking. J. Shanxi Univ. Fin. Econom. **13**(3), 74–84 (2017)
3. G. Hilary, K.W. Hui, Does religion matter in corporate decision making in America? J. Financ. Econom. **93**(3), 455–473 (2009)
4. S.T. McGuire, T.C. Omer, N.Y. Sharp, The impact of religion on financial reporting irregularities. Account. Rev. **87**(2), 645–673 (2012)
5. S.D. Dyreng, W.J. Mayew, C.D. Williams, Religious social norms and corporate financial reporting. J. Bus. Finance Account., **39**(39), 845–875 (2012)
6. P.J. Zak, S. Knack, Trust and growth. Econom. J. **470**(111), 295–321 (2001)
7. L. Guiso, P. Sapienza, L. Zingales, Trustin the stock market. J. Fin. **63**(6), 2557–2600 (2008)
8. T. Opler, L. Pinkowitz, R. Stulz, R. Williamson, The determinants and implications of corporate cash holdings. J. Financ. Econ. **52**(1), 3–46 (1999)
9. L. Yujun, L. Xingyun, S. Zhi, Industrial features of firms' cash holding: difference and convergence. Account. Res. **285**(7), 66–72 (2011)
10. C. Deqiu, L. Sifei, W. Cong, Government quality, ultimate property rights, and corporate cash holdings. Manage. World **218**(11), 127–141 (2011)
11. Y. Xin, L. Xinchun, X. Liping, Local religious tradition and the sources of start-up capital of private firms. Econ. Res. J. **79**(4), 161–173 (2016)
12. Y. Dezhu, H. Mengxi, Religious tradition, rule of law and corporate risk taking. Res. Financ. **402**(5), 95–103
13. Z. Weiying, K. Rongzhu, Trust in china: a cross-regional analysis. Econ. Res. J. **412**(10), 59–70 (2002)

14. L. Jinbo, Internal control, social trust, and the production efficiency. J. Account. Econ. **108**(3), 72–91 (2017)
15. H. Jingtong, Fan Ruoxuan, Social trust and firm cash holdings: an explanation based on trade-off theory. J. Shanghai Univ. Finance Econ. **96**(4), 30–41 (2015)
16. H. Zhiying, Female CEO, Social trust and financial constraints. Econom. Manage. **557**(8), 88–98 (2015)
17. C. Kim, D. Mauer, A. Sherman, The determinants of corporate liquidity: theory and evidence. J. Finan. Quant. Anal. **33**(3), 305–334 (1998)
18. J. Harford, S.A. Mansi, W.F. Maxwell, Corporate governance and firm cash holdings in the U.S. Soc. Sci. Electron. Publishing **87**, 535–555 (2012)
19. S. Yupeng, L. Xinrong, Common resources and social trust: evidence from mandatory education experience in China. Econ. Res. J. **575**(5), 86–100 (2016)
20. W. Xiaolu, F. Gang, *China's Marketization Index by Provinces (CMI 2016* (Social Science Literature Press, Beijing, 2017)

Data-Driven Organizational Structure Optimization: Variable-Scale Clustering

Ai Wang and Xuedong Gao

Abstract With the continuous improvement of external data acquisition ability and computing power, data-driven optimization of organizational structure becomes an emerging technique for various enterprises to develop business performance and control management costs. This paper focuses on the management scale level discovery problem for the optimization of enterprise organizational structure. Firstly, according to the scale transformation theory, the scale level of the multi-scale dataset is defined. Then, a scale level discovery method based on the variable-scale clustering (SLD-VSC) is proposed. After determining management objectives, the SLD-VSC is able to recognize optimal management scale level and the scale characteristics of each management object clusters distributed in different management scale levels. The numerical experimental results illustrate that the proposed SLD-VSC is able to support enterprises improving their organizational structure by identifying the management scale levels from business data.

Keywords Variable-scale clustering · Scale transformation · Management scale level · Organizational structure

1 Introduction

The organizational structure always has a significant impact on various types of enterprises in developing business performance and controlling management cost. Common enterprise organizational structures (like management layers and business departments) are mostly founded artificially following the subjective experience of leadership group. That always suffers from the frequent business changes caused by market fluctuation.

A. Wang · X. Gao (🖂)
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: gaoxuedong@manage.ustb.edu.cn

A. Wang
e-mail: wangai22222@126.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_6

With the continuous improvement of data acquisition ability and computing power, enterprises now have the potential to predict their internal business variation tendency by mining external market environment information, and then adjust their organizational structure for clear management objectives, so as to achieve the data-driven organizational structure optimization.

Compared with the traditional organizational structure establishment, the advantages of the data-driven organizational structure establishments are as follows. First, ensure all management layers have clear management objectives and tasks. On the one hand, for the current management hierarchy with unclear management objective (that is redundant organizational structure), the functions of these management layer and related business layer could be merged together to reduce management cost. On the other hand, for the clear management objective levels without any matched organizational structure (that is absent organizational structure), flexible organizational structure (such as project teams) could be supplemented to expand enterprise business scope. Secondly, respond to dynamic market changes in time. The establishment of enterprise organization transforms from leaders' decision-making ability dependent to objective data feature dependent (including external market environment data and internal business status data), which improves the scientific of enterprise organization management, as well as enterprise competitiveness.

What stands the most to realize data-driven optimization of organizational structure is to clarify the data demand characteristics between personnel at different management levels. Since the main business served by personnel of different management levels is quite the same, only the business scope they are responsible for is different. It can be seen that the data demand of personnel at different management levels are characterized by the same observation dimension but with different observation scales. Therefore, multi-scale data analysis methods should be utilized to study the optimization of enterprise organizational structure.

This paper studies the data-driven organization structure optimization problem based on the scale transformation theory. The main contributions are as follows. Firstly, we analyze the advantages of the scale transformation theory, that could enrich the existing theoretical system of big data analysis by establishing the de-scenario, de-human and de-label variable-scale data analysis mechanism. Secondly, after defining the scale level of the multi-scale dataset, a scale level discovery method based on the variable-scale clustering (SLD-VSC) is proposed. In order to identify the optimal scale level, two scale transformation strategies are utilized to the management scale level discovery process. Finally, we test the accuracy and efficiency of the proposed SLD-VSC in the context of customer segmentation using four classic evaluation metrics. Compared to the traditional single-scale clustering algorithm, the SLD-VSC could obtain more accurate clustering results under high efficiency. And the experimental results also illustrate that the SLD-VSC could directly support enterprises improve their organizational structure by identifying the optimal management scale levels.

2 Scale Transformation Theory

Under the data economy circumstances, big data analysis technology has paid more attention on the integrated development of multi-industry scenarios, not the traditional data mining methodology. The big data supported intelligent technology (such as manufacturing intelligence and business intelligence), even has developed various data analysis methods and software platform tools with independent industry characteristics. Although that decentralized scenarios (applications) driven mode could continuously stimulate the research passion of big data analysis field, an automatic data analysis mechanism and method that could support the demand of all business scenarios (de-scenario) is still missing in the existing theoretical system of big data analysis.

Scale is a philosophical concept that exists simultaneously in both humanoid decision making (intelligent decision making) science and natural science. The problem solving theory in cognitive psychology considers that decision making is a thinking process, in which people construct a problem solving space from multiple scales and obtain the final satisfied solution through changing observation scales of the problem. Experiments in different natural science fields find the same results that the scale effect is a widely existing objective phenomenon that the nature or characteristics of an object will change following the transformation of its observation scale (such as time scale, space scale, etc.).

The scale transformation theory (ST) mainly studies the scale representation and transformation problem in data analysis process, which is the first to introduce the observation scale concept and scale transformation feature to big data analysis. The scale transformation theory could not only create the direct connection between subjective preferences and objective data using the hierarchical observation scale model, but also simulate data analysts' thinking activity within their decision making process using scale transformation mechanism. The intelligent scale computing framework, that acts as one of the philosophy computing techniques, plays a significance role in the scale transformation theory, which is used to realize the automation of the whole data analysis process.

Several researchers have started to establish the mechanism and methods of the variable-scale data analysis based on the ST. Gao [1, 2] considers that the observation scale in data analysis process consists of two components: one is all possible dimensions to describe objects, the other is all possible values that an object can get on every dimension. Scale itself is a kind of objective priori knowledge, which is not affected by any observed object. There are different types of scales distinguished by the purpose and characteristics when describing objects. The basic scale refers to the standard that is unable to be further divided when describing objects (such as time scale, space scale, etc.). While the observation scale refers to all the scales utilized to describe an specific object in the same space-time. In addition, Wang [3–5] built the concept space model (CS) based on the basic concepts of observation scales, in order to describe the hierarchical structure relationship between different observation scales. In the CS, the first component of scale is represented by the concept chain, and the other component is represented by the value space.

According to the cross-industry standard process for data mining (CRISP-DM) [6], six different activities (that is business understanding, data understanding, data preparation, modeling, evaluation and deployment) could be divided into three phases following task objectives.

The first phase consists of two activities (i.e., business understanding and data understanding), which aims to identify data mining tasks [7, 8]. In the beginning, managers describe the actual requirements to data analysts, then data analysts complete the data detection work, and finally jointly determine the data analysis themes and tasks with the business leader [9, 10].

The second phase consists of three activities (i.e., data preparation, modeling and evaluation), which aims to obtain data mining results [11]. In this stage, data analysts firstly select the initial dataset for each task following their experience, then match the models or algorithms to obtain the analysis results, and finally manager evaluates those results. If the actual demands are unable to be solved and satisfied appropriately, the analysis theme and tasks will be re-planned through returning to the first phase.

The third phase consists of one activity (i.e., deployment), which aims to apply data mining results to actual business scenarios. In this stage, manager verifies the effectiveness and performance of the data analysis results in practice, and timely push feedbacks or requirements to data analysts for planning the new data analysis work [12, 13].

It can be seen that the CRISP-DM is a multiple iteration process with high communication costs between manager and data analysts. Besides, data analysts always play the intermediary role between actual users (the demand-side enterprise represented by the manager) and users' own business data. The quality of analysis results are heavily affected by data analysts' experience, which leads to the subjective decision-making risks. Facing the (de-human) development trend of big data analysis technology, the variable-scale data analysis could simulate data analysts' thinking and decision-making activities via the scale transformation mechanism [14].

Moreover, lots of evidences show that it is large amounts of high-quality labeled data that have fully improved the accuracy of original models and algorithms, and thus rapidly promoted the application of big data analysis technology in recent periods. Especially for supervised learning methods in the machine learning field, the influence of data labels on analysis results could even exceed the training effect brought by optimizing algorithm steps [15].

Although the current relatively mature data acquisition and storage techniques are able to ensure the continuous collection of more (manually) labeled high-quality datasets, this training oriented data analysis mode limits the analysis results will always lag behind the speed at which people generate their new knowledge and experience. Facing the (de-label) development trend of big data analysis technology, the variable-scale data analysis could autonomously learn and infer information and knowledge from raw data via the intelligent scale computing framework.

3 Management Scale Level Discovery

3.1 Management Scale Level

In the scale transformation theory, a multi-scale dataset is expressed by $D^S = \left(U, A^S, V^S, f\right)$, where U is the object set; $A^S = \left\{A^1, A^2, \ldots, A^r\right\}$ is the attribute set, and there is at least one attribute that has multiple scales, i.e., $\exists A^\lambda, A^\lambda = \left\{A_0^\lambda, A_1^\lambda, \ldots, A_n^\lambda\right\}(\lambda = 1, 2, \ldots, r)$; V^S is the object value domain following the information function $f : U \times A^S \rightarrow V^S$.

Therefore, the scale level (SL) of D^S can be defined as $SL = \left\{A_i^\lambda\right\}$, where $|SL| = r$ and $\left|A^\lambda \cap SL\right| = 1$ $(\lambda = 1, 2, \ldots, r)$. Let n^λ represents the total number of scales in attribute A^λ, it can be seen that D^S could generate $\prod_{\lambda=1}^{r} n^\lambda$ different scale levels.

The scale transformation strategy (STS) is utilized to identify which scale level is qualified or satisfied for the variable-scale data analysis. Since data analysts have the obvious behavioral preference in decision-making during the CRISP-DM process, the original scale transformation strategy is classified into two types, i.e., the optimistic scale transformation strategy (OSTS) and pessimistic scale transformation strategy (PSTS).

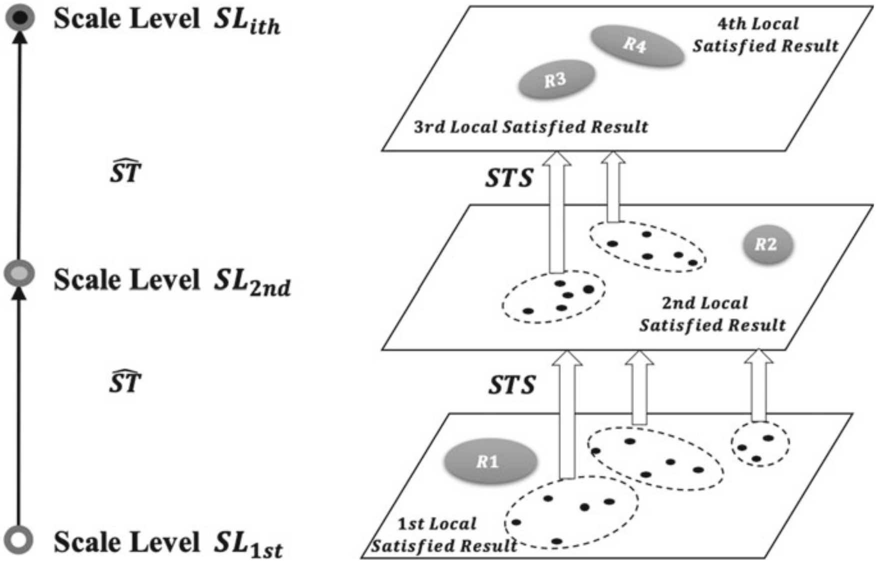


Fig. 1 Example: the scale up transformation process $(\widehat{ST})$

3.2 Scale Level Discovery Based on the Scale Transformation

According to the scale transformation process (see Fig. 1), the scale level discovery method based on the variable-scale clustering (SLD-VSC) is proposed. The algorithm steps are as follows.

Algorithm 1 *Scale Level Discovery Algorithm Based on the Variable-Scale Clustering (SLD-VSC)*

```
Input: Single-scale initial dataset $D^0$, Concept space $CS$, number of clusters $k$, Management objective metric $\xi$;
Output: Satisfied clusters $R^S$ with scale characteristics under optimal scale level $SL^S$.
Step 1: Take the scale level of $D^0$ as the basic scale level $SL^0$ , and conduct the initial clustering on $D^0$ using $k$.
Step 2: Evaluate the initial clustering result $R^0$ using $\xi$, and set the current results as the optimal results, that is $R^S = R^0$ and $SL^S = SL^0$.
Step 3: Establish the multi-scale dataset $D^S$ for the initial dataset $D^0$ using the $CS$;
Step 4: Scale up transformation.
Step 4.1: If the OSTS is adopt, improve the scale of the attribute that has the highest scale transformation value in $D^S$, and establish the transformed single-scale dataset $D^T$.
Step 4.2: If the PSTS is adopt, improve the scale of the attribute that has the lowest scale transformation value in $D^S$, and establish the transformed single-scale dataset $D^T$.
Step 5: Conduct the clustering analysis on $D^T$, and evaluate the clustering results $R^T$ using $\xi$.
Step 6: Evaluate scale transformation effect.
Step 6.1: If the current results obtain better performance on $\xi$ than the previous optimal results, then update the satisfied results by $R^S = R^T$ and $SL^S = SL^T$, and go to Step 4.
Step 6.2: If the current results obtain worse performance on $\xi$ than the previous optimal results, then output the optimal results $R^S$ and $SL^S$.
```

The time complexity of the SLD-VSC is $O(tn^r)$, where r is the number of attributes, n is the maximum number of observation scales that one attribute contains in the multi-scale dataset D^S, and t is once (meta) clustering time.

4 Experiment Results and Discussion

In this section, a numerical experiment is designed to verify the accuracy and efficiency of the proposed SLD-VSC method. The experimental purpose is to divide customers into different clusters with clear management scale levels, which could support enterprise to hire the appropriate number of customer mangers and design differential marketing strategies.

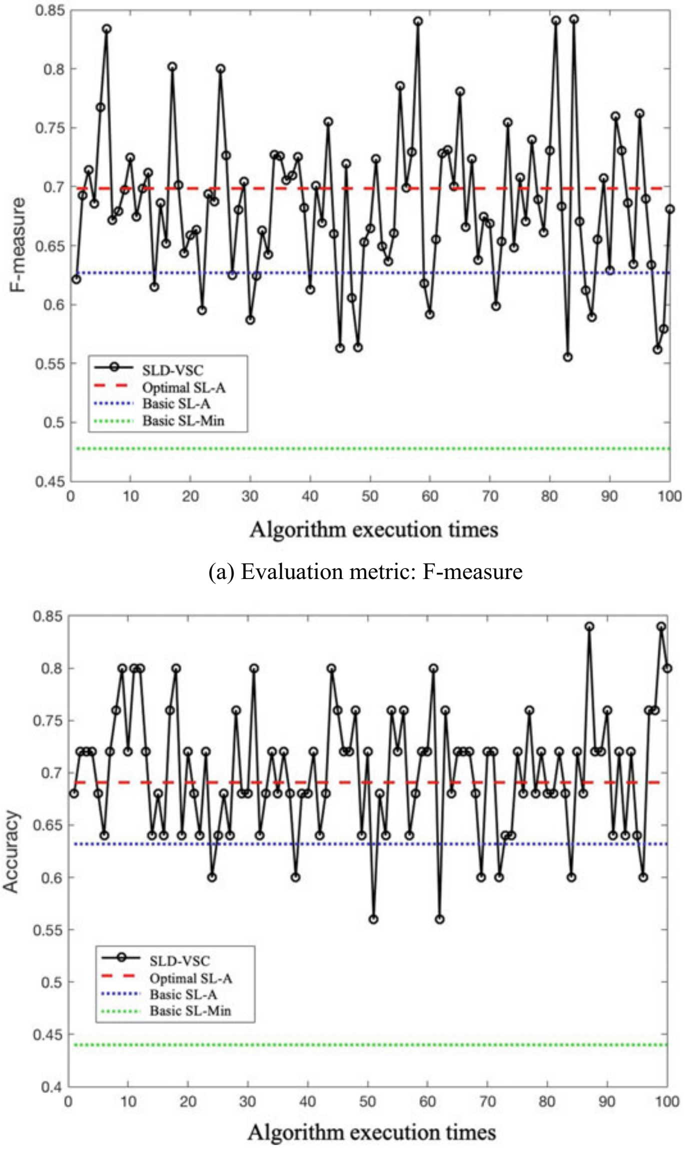
Table 1 shows the data preparation result D^S of a single-scale customer dataset, which includes twenty five customers U, four customer attributes $A^S = \left\{C^1, C^2, C^3, C^4\right\}$, and one management objective attribute d.

According to the scale level definition in Sect. 3, the basic management scale level of the multi-scale customer dataset D^S is $\left\{C_0^1, C_0^2, C_0^3, C_0^4\right\}$. After calculating all the scale levels generated from D^S using the traditional single-scale clustering method k-modes, the optimal scale level is $\left\{C_1^1, C_1^2, C_1^3, C_0^4\right\}$ (see the grey area of Table 1).

Figure 2 shows the customer segmentation experimental results. On the one hand, four classic metrics (i.e., F-measure, Accuracy, NMI and RI) are utilized to evaluate the accuracy of the SLD-VSC. On the other hand, experiments repeat a hundred

Table 1 The multi-scale customer dataset

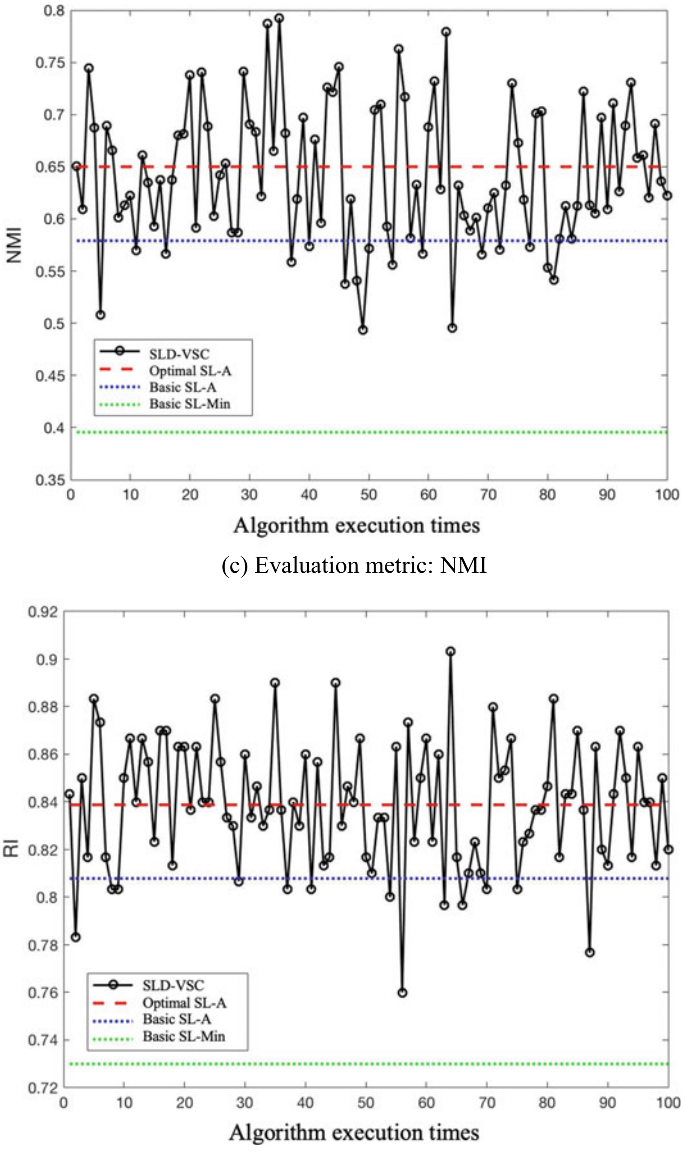
U	C_0^1	C_1^1	C_2^1	C_0^2	C_1^2	C_2^2	C_0^3	C_1^3	C_2^3	C_0^4	C_1^4	D
x_1	a	B	1	a	A	0	a	D	1	a	1	+
x_2	c	D	1	c	A	0	c	C	2	b	0	*
x_3	b	B	1	b	A	0	c	C	2	a	1	+
x_4	d	D	1	e	C	0	d	A	2	a	1	/
x_5	c	D	1	d	A	0	c	C	2	a	1	*
x_6	e	C	0	f	C	0	b	E	1	b	0	#
x_7	d	D	1	e	C	0	c	C	2	b	0	*
x_8	g	A	0	g	B	1	e	B	0	c	0	\
x_9	f	A	0	f	C	0	e	B	0	a	1	/
x_{10}	g	A	0	f	B	1	f	B	0	d	0	\
x_{11}	g	A	0	f	C	0	d	A	2	c	0	\
x_{12}	f	A	0	f	C	0	a	D	1	b	0	#
x_{13}	e	C	0	f	C	0	d	A	2	a	1	#
x_{14}	a	B	1	a	A	0	a	D	1	a	1	+
x_{15}	b	B	1	c	A	0	c	C	2	a	1	*
x_{16}	c	D	1	d	A	0	b	E	1	b	0	–
x_{17}	e	C	0	e	C	0	d	A	2	b	0	#
x_{18}	g	A	0	f	B	1	f	B	0	d	0	\
x_{19}	g	A	0	f	B	1	e	B	0	d	0	\
x_{20}	c	D	1	e	C	0	b	E	1	a	1	/
x_{21}	d	D	1	e	C	0	a	D	1	a	1	/
x_{22}	b	B	1	b	A	0	b	E	1	b	0	–
x_{23}	a	B	1	b	A	0	b	E	1	a	1	+
x_{24}	e	C	0	f	C	0	c	C	2	a	1	#
x_{25}	f	A	0	g	B	1	e	B	0	b	0	\



(a) Evaluation metric: F-measure

(b) Evaluation metric: Accuracy

Fig. 2 The experimental results of the SLD-VSC



(c) Evaluation metric: NMI

(d) Evaluation metric: RI

Fig. 2 (continued)

times under each metric to verify the efficiency of the SLD-VSC (see the solid line of Fig. 2). In addition, three evaluation standard are precalculated by the single-scale clustering algorithm k-modes, i.e., the minimum external validity under the basic scale level (Basic SL-Min), the average external validity under the basic scale level (Basic SL-A) and the average external validity under the optimal scale level (Optimal SL-A), which are shown in three horizontal lines of Fig. 2 respectively from the lowest level to the highest.

Since the solid line keeps being above the lowest horizontal line, the experimental results illustrate that the accuracy of the SLD-VSC is higher than the lowest performance of the traditional clustering algorithm. Since the solid line is above the middle horizontal line in most cases, and sometime even above the highest horizontal line, which verifies the high efficiency of the SLD-VSC.

5 Conclusions

With the continuous improvement of external data acquisition capability and computing capability, the establishment of data-driven organizational structure optimization technology supports enterprises sustainably enhancing their competitiveness. This paper studies the management scale level discovery problem for optimizing the organizational structure of enterprises. After defining the scale level of the multi-scale dataset, a scale level discovery method based on the variable-scale clustering (SLD-VSC) is proposed. We test the accuracy and efficiency of the SLD-VSC in the context of customer segmentation. The numerical experimental results illustrate that the SLD-VSC could directly support enterprises improve their organizational structure by identifying the optimal management scale level.

Enterprises could apply the SLD-VSC to assess whether their current management layers all have clear management objectives using market real-time data, so as to timely reduce redundant organizational structures or add absent organizational structures.

References

1. X. Gao, A. Wang, Variable-scale clustering, in *Proceeding of the 8th International Conference on Logistics, Informatics and Service Sciences* (Toronto, Canada, 2018), pp. 221–225
2. X. Gao, A. Wang, Customer satisfaction analysis and management method based on enterprise network public opinion, in *Operations Research and Management Science* (In Press, 2019)
3. A. Wang, X. Gao, Multifunctional product marketing using social media based on the variable-scale clustering. Tech. Gaz. **26**(1), 193–200 (2019)
4. A. Wang, X. Gao, Hybrid variable-scale clustering method for social media marketing on user generated instant music video. Tech. Gaz. **26**(3), 771–777 (2019)
5. A. Wang, X. Gao, M. Yang, Variable-scale clustering based on the numerical concept space, in *Proceeding of the 9th International Conference on Logistics, Informatics and Service Sciences* (Maryland. US) (2019), pp. 65–69

6. S. Wu, X. Gao, M.M. Bastien, *Data Warehousing and Data Mining* (Metallurgical Industry Press, China, 2003), pp. 148–155
7. A. Wang, X. Gao, Technique of data mining tasks discovery for data mining, in *Proceeding of the 7th International Conference on Logistics, Informatics and Service Sciences, Kyoto. Japan* (In Press, 2017)
8. A. Wang, X. Gao, Multi-tasks discovery method based on the concept network for data mining. *IEEE Access, 7*, 139537–139547 (2019)
9. X. Chen, *Technology of Thinking Processes Discovery for Data Mining Application* (University of Science and Technology Beijing, Beijing, 2012)
10. K. Gu, *Technology of Concept Pair Identification for Thinking Theme Discovery* (University of Science and Technology Beijing, Beijing, 2013)
11. A. Wang, X. Gao, M. Tang, Computer supported data-driven decisions for service personalization: a variable-scale clustering method. Stud. Inf. Cont. **29**(1), 55–65 (2020)
12. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik-Tech Gaz **25**(2), 528–535 (2018)
13. A. Wang, X. Gao, Intelligent computing: knowledge acquisition method based on the management scale transformation. Comput. J. (2020) (In Press)
14. J. Li, S.X. Pan, L. Huang, X. Zhu, A machine learning based method for customer behavior prediction. Tehnicki Vjesnik-Tech. Gaz. **26**(6), 1670–1676 (2019)
15. L.M. Wang, Z.Y. Hao, X.M. Han, R.H. Zhou, Gravity theory-based affinity propagation clustering algorithm and its applications. Tehnicki vjesnik-Tech. Gaz. **25**(4), 1125–1135 (2018)

Auto-carrier Transport Routing Problem with Loading Constraint

Xiqing Yang, Wenliang Bian, Hanping Hou, and Xiying Yang

Abstract Vehicle logistics refers to the whole process of rapid delivery of vehicles distributed by logistics company according to customer orders. Auto-carriers are used as the transportation tool, which efficiency and loading rate directly determine the logistics cost. Compared with traditional vehicle routing problems, route planning of auto-carriers also needs to consider how to load the vehicles, which has higher requirements for the model formulation. In this paper we provide a three-dimensional loading constraint and build a multi-objective mixed integer programming model based on the minimum total cost, the minimum number of auto-carriers and the shortest total driving distance. We use Gurobi as a solver to pre-process multiple constraints and solve the model. The experimental results show that the model can find the optimal solution efficiently and the algorithm can give an accurate solution in a limited time, which indicates that the method designed in this paper is feasible and even optimal for certain problem scales.

Keywords Auto-carrier transportation · Routing problem · Loading constraint · Mixed integer programming

X. Yang (✉) · W. Bian · H. Hou · X. Yang
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 19120591@bjtu.edu.cn

W. Bian
e-mail: wlbian@bjtu.edu.cn

H. Hou
e-mail: hphou@bjtu.edu.cn

X. Yang
e-mail: 19120592@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_7

1 Introduction

With the economic development, the demand for vehicles in China has increased significantly, reached 20 million in 2019. However, China is a vast country so that the production and sales places are often far away. Vehicles manufacturers need to order logistics companies to fulfill transportation tasks across the country according to car purchase orders from customers all over the country [1]. The logistics company formulates the corresponding distribution plan, and pursues the optimal transportation cost while ensuring the completion of the task. However, due to the high cost and small number of vehicles, the transportation is limited by the length, height, and weight of the vehicles. At present, the utilization rate of the transportation space of the auto-carrier is low and the industry economy is poor. Despite its practical significance, there are not many theoretical studies on vehicle logistics.

Agbegha et al. [2] were the first to raise the issue of car transport. They describe how companies transport cars to dealers in the US market, and then focus on loading sub-problems. Tadei [3] aims at maximizing the profit of the vehicles transport company, and considers the loading scheme and transportation route of the car at the same time, but Tadei only give an approximate solution to the problem. Bonassa [4] proposes a mixed integer programming method to solve the dynamic multi-period auto-carrier transportation problem. He considers the transportation cost of the real world automobile transportation company, the calculation results show that the application of the mathematical model greatly reduces the cost saving rate. But he did not consider the waste of transportation costs caused by the sub-loop. Since its development, the research trend of this problem is as follows: (1) With the background of auto-carrier transportation, only one of the sub-problems such as vehicle scheduling, routing optimization, and loading decision-making is studied. (2) It is regarded as a traditional path planning problem, adding time window and other practical constraints, and a heuristic algorithm is designed to solve it. With the expansion of the entire vehicle market, the issue of auto-carrier transportation should be studied as a whole, so as to help logistics companies achieve optimal decisions and improve competitiveness.

Therefore, the research in this paper is very challenging because it combines three NP-hard problems: loading, vehicle selection, and routing optimization. First, a multi-objective hybrid overall planning model with three-dimensional loading constraints is established, and Gurobi is used to solve the overall problem. The numerical results show the effectiveness and superiority of the model and method: it can output accurate stowage schemes according to not only the type of vehicles but also the loading limits of vehicles to provide the optimal path. Meanwhile, common sub loop problems can be solved efficiently. Most literature based on the goal of minimizing the transportation gross cost to build the model. In this research, the hierarchical weighting method in Gurobi is used to achieve multi-objective optimization, which is more in line with actual needs.

2 Problem Description

Auto-carrier Transportation Problem (ATP) is a special type of path planning problem. The background is how vehicle logistics companies can complete vehicle distribution plans quickly and efficiently according to customer orders, thereby minimizing logistics costs. For the case of delivering multiple orders, considering the route planning problem of multiple auto-carriers load multiple vehicles. To solve this problem, three aspects need to be considered: the best loading solution for the auto-carriers; the type of auto-carriers selected; 3.the optimal driving path of the auto-carriers.

These three aspects affect each other and make the problem very complicated. In the solution process, it is also necessary to consider the computational scale and the operational limitations such as the existence of sub-loops.

In the domestic vehicle logistics market, there are usually three types of auto-carrier, which are type 1-1: the upper and lower layers are in a row; type 1-2: the lower layer is in a row and the upper layer is in two rows; type 2-2: the upper and lower layers are two rows. An example of a type 1-1 carrying five vehicles [5] is depicted in Fig. 1.

As shown in the figure below, there are different models of auto-carriers to adapt to different loading vehicles. If they are used improperly, it will cause bumps in the transportation process, resulting in safety hazards. The correct placement method should be as shown in Fig. 2. Therefore, how to match the size of the auto-carriers and the loading vehicle, this brings the first part of our research problem. UPS, as a 3PL enterprise that provides vehicle transportation, it serves multiple vehicle manufacturers, so there are many different types and unspecific loading vehicles to be shipped. Therefore, in the whole vehicle logistics, we need to consider the joint optimization of loading and travel-routing problem.

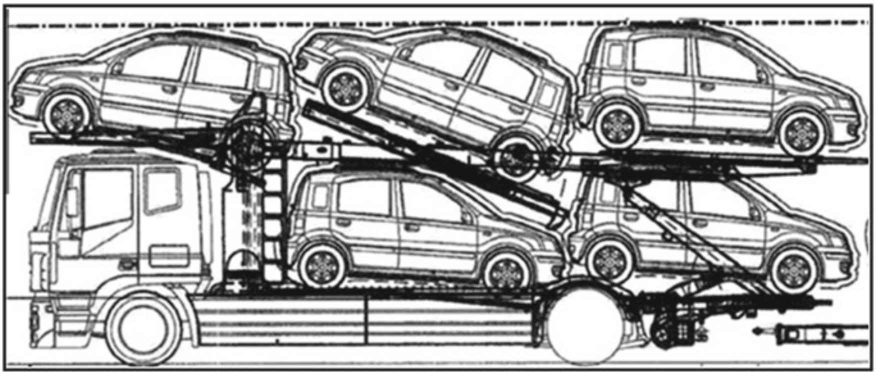


Fig. 1 An example of an auto-carrier 1-1 carrying five vehicles

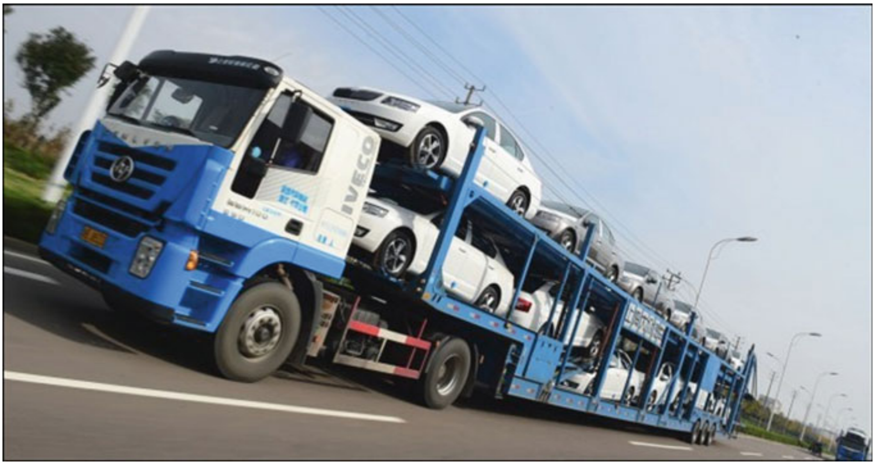


Fig. 2 A correct loading after optimized settings

The problem of loading vehicles into an auto-carrier can be considered a specific three-dimensional loading problem with a considerable number of complex constraints. Since the loading capacity of an auto-carrier is mainly depends on the length, width and height of the vehicles, based on the actual situation, we have simplified the following loading requirements.

- The upper and lower loading areas of each auto-carrier can be regarded as rectangles equivalently, and the vehicles in each row are placed vertically.
- The total length of each row of vehicles and the safety distance of each row does not exceed the length of the auto-carrier.
- Maintain a safe distances in the longitudinal and transverse directions of adjacent vehicles.
- Type 1-1 auto-carriers are loaded with one row of vehicles, type 1-2 auto-carriers are loaded with two rows on the upper layer and one row of vehicles are loaded on the lower layer, and two rows of vehicles are mounted on the upper and lower floors of the type 2-2 auto-carriers.
- Vehicles with a height exceeding 1.7 meters can only be installed on the lower floors of type 1-1 and 1-2.
- The lower layer should be filled as much as possible, and the upper two columns should be as symmetrical as possible.

In the whole auto-carrier transportation process, the first impact on the cost is the number of auto-carriers used. Second, the use of auto-carriers with the same number of types, the use of auto-carriers costs reduced. In the same type of auto-carrier, the use cost of type 1-1 is lower, type 2-2 is higher, and type 1-2 is slightly lower than the average of the former two.

$$\cos t_{1-1} < \cos t_{1-2} < \frac{\cos t_{1-1} + \cos t_{2-2}}{2} < \cos t_{2-2} \tag{1}$$

Third, when the number of auto-carriers and their types are all equal, the shorter the mileage, the lower the cost. Therefore, in order to maximize the benefits, we will formulate mathematical models for the following three optimization goals.

- Minimize transportation costs
- Minimize the travel distance
- Minimize the number of auto-carriers used.

3 Mathematical Model Establishment

ATP is an overall planning problem under complex constraints. The complete vehicle logistics loading and route selection problems with loading constraints that minimize the total transportation distance and minimize costs can be formulated as a mixed integer programming model for multi-objective programming. To build a mathematical model, we first define the following symbols and variables:

3.1 Parameters

$N = \{0, 1, 2, \ldots, n, n+1\}$, represents the set of different transportation nodes, '0' represents the starting point of transportation spot;

$K = \{1, 2, \ldots, k\}$, represents the set of various auto-carriers;
$V = \{1, 2, \ldots, n\}$, represents the set of vehicles to be delivered to dealer;
c_k: the usage cost of auto-carrier k;
F_k: the usable loading area number of auto-carrier k;
L_k^f: the length of area f of auto-carrier k;
W_k^f: the width of area f of auto-carrier k;
H_k^f: the height of area f of auto-carrier k;
s: the safety distance between any kind of vehicles;
l_v: the length of vehicle v;
w_v: the width of vehicle v;
h_v: the height of vehicle v;
d_{ij}: the distance from the delivery point i to the demand point j.

3.2 Variables

x_{ij}^k $binary$, represents whether the auto-carrier k through the point i to point j;

y_v^{kf} $binary$, represents whether the vehicles are loaded in the f area of the auto-carrier k;

z^k $binary$, represents whether the car k is used.

3.3 Objectives

The objective functions are constructed as below according to the hypothesis and analysis.

$$Minimize\,obj1 = \sum_{k \in K} c_k z^k \tag{2}$$

$$Minimize\,obj2 = \sum_{k \in K} \sum_{i \in N} \sum_{j \in N} d_{ij} x_{ij}^k \tag{3}$$

$$Minimizeobj3 = \sum_{k \in K} z^k \tag{4}$$

Objective function (2) minimizes the transportation costs of all auto-carriers. Objective function (3) minimizes the total passing distance of all auto-carriers. Objective function (4) minimizes the total number of all auto-carriers. When solving the objective function, the gurobi software can be used to weight it and achieve the optimization goal in order. According to the actual situation, the sorting method is: cost → route → number of auto-carriers.

3.4 Constraints

Constraint (5) represents that each vehicle must be allocated:

$$\sum_{j \in N/\{0\}} \sum_{k \in K} x_{ij}^k = 1, \forall i \in V \tag{5}$$

Constraint (6) represents the auto-carrier must pass a certain point to load the vehicles corresponding to that point:

$$\sum_{f \in F_k} y_i^{kf} \geq x_{ij}^k,\ \forall i \in V, j \in N/\{0\}, k \in K \tag{6}$$

Constraint (7) is used to make path balance constraints:

$$\sum_{i \in N/j} x_{ij}^{k} - \sum_{i \in N/j} x_{ij}^{k} = \begin{cases} -1\ j = 0 \\ 0\ j \in V, \forall k \in K \\ 1\ j = n + 1 \end{cases} \tag{7}$$

Constraint (8) indicates whether the auto-carrier is used:

$$y_i^{kf} \le z^k, \ \forall i \in V, k \in K, f \in F_K \tag{8}$$

Constraints (9) and (10) represent loading limits for the length and width of the vehicles, respectively.

$$\sum_{i \in V} (s + l_i) y_i^{kf} \le L_k^f, \ \forall k \in K, f \in F_K \tag{9}$$

$$(s + w_i) y_i^{kf} \le W_k^f, \ \forall k \in K, f \in F_k \tag{10}$$

In addition to the above constraints, we also consider the constraints-height, which will be implemented using the preprocessing of judgment in Gurobi. Moreover, in order to solve the situation during transportation that the sub-loops cause waste of transportation resources, we have added a sub-loop constraint. It will add lazycut to the callback function for implementation.

4 Algorithm Design

4.1 Gurobi Solver

The mathematical model belongs to MIP (Mixed Integer Programming Model). Researchers usually use Lingo software to solve comparatively small integer programming models. However, this model has relatively many variables and constraints, which is difficult for Lingo to calculate the results in a limited time. In this paper, we use Gurobi optimization software in Python to solve this problem. Gurobi is a new generation of large-scale mathematical planning optimizer developed by American Gurobi company. In the third-party optimizer evaluation held on the Decision Tree for Optimization Software website, Gurobi showed faster optimization speed and accuracy. It is currently one of the world's top software packages for solving linear programming, integer programming, and some non-linear programming, which is widely used in many industries.

4.2 Floyd Algorithm

Floyd algorithm, also called interpolation method, is an algorithm for finding the shortest path between multiple source points in a given weighted graph. Floyd algorithm is used to solves the shortest path between any two points, which is also used to calculate the transitive closure of a directed graph. This algorithm is simple and effective, and because of its compact triple loop structure, the planning efficiency for dense graphs is higher than that of Dijkstra's algorithm. According to the transportation characteristics of vehicle logistics, we select the Floyd algorithm as the path planning algorithm.

The Floyd algorithm is mainly used to find the shortest path with multiple sources and no negative weight edges. It use matrix to record point-to-distance map. Time complexity of Floyd algorithm is $O(V^3)$ and space complexity is $O(V^3)$. The formula for the Floyd algorithm is shown in Eq. (11).

$$G[s][g] = \min\{G[s][g], G[s][i] + G[i][g]\}(1 < i < n) \tag{11}$$

The calculation process of the Floyd algorithm is as follows. Start from any unilateral path. The distance between all two points is the weight of the edge. If no edge is connected between the two points, the weight is infinite. For each pair of vertices i and j, see if there is a vertex k that makes the path from i to k to j shorter than the known path. If it is, update it.

We use the adjacency matrix G to represent the connection graph of point pairs, if there is a path reachable from V_i to V_j, then $G[i, j] = d$ represents the length of the path. Otherwise, $G[i, j] = \infty$. A matrix V_p is defined to record the information of the inserted points, and to compare the distance after the insertion point is the same as the original distance (12).

$$G[i][j] = \min\{G[i][j], G[i][k] + G[k][j]\} \tag{12}$$

If the value of $G[i, j]$ becomes smaller, V_k is added to the path matrix V_p.

4.3 Solving Process

We first establish the optimal loading scheme model with the smallest number of auto-carriers, and use this scheme to distribute vehicles until the transportation task is completed to obtain the initial number of auto-carriers. Secondly, based on the goal of minimize the cost of auto-carriers, we design the quantity adjustment model to optimize the initial number of auto-carriers and obtain the optimal loading scheme model. Based on the above optimization model, we calculate the minimum number of auto-carriers and establish the shortest mileage model according to the limited number of auto-carriers. Then the loading scheme with the smallest total mileage is selected as the optimal distribution scheme. Finally, we establish a path-based logistics transportation loading model, and use the Floyd algorithm to calculate the shortest path between any nodes. A global search algorithm is designed to obtain the reasonable distribution scheme, and to eliminate the sub-loops as the goal to optimize the scheme. According to the above ideas, the improved algorithm is designed. The following are the flow table and some important codes of our algorithm.

(1) *Build a model*

```
def subtourelim(model, where):
    if where == GRB.Callback.MIPSOL:
        vals = model.cbGetSolution(model._vars)
         selected = {}
         for i in model._vars.keys():
             if vals[i] > 0.9:
                 if i[2] not in selected.keys():
                     selected[i[2]] = [(i[0],i[1])]
                 else:
                      selected[i[2]].append((i[0],i[1]))

 for k in selected.keys():
     temp = tuplelist(selected[k])
     tour = subtour(temp)
     for cy in tour.keys():
         if len(tour[cy]) < len(selected[k]):
             expr = LinExpr()
             for i in range(len(tour[cy])):
                 edges = temp.select(tour[cy][i],'*')
        if len(edges)!=0:
        expr += model._vars[edges[0][0], edges[0][1], k]
  model.cbLazy(expr <= len(tour[cy])-1)
```

(2) *Set muti-objectives*

```
model.setObjectiveN(z.sum(‘*’),index = 0,priority = 3,name = ’obj1’)
    model.setObjectiveN(x.prod(xindex),index = 2,priority = 2, name = ’obj3’)
    model.setObjectiveN(z.prod(zindex),index = 1,priority = 1, name = ’obj2’)
```

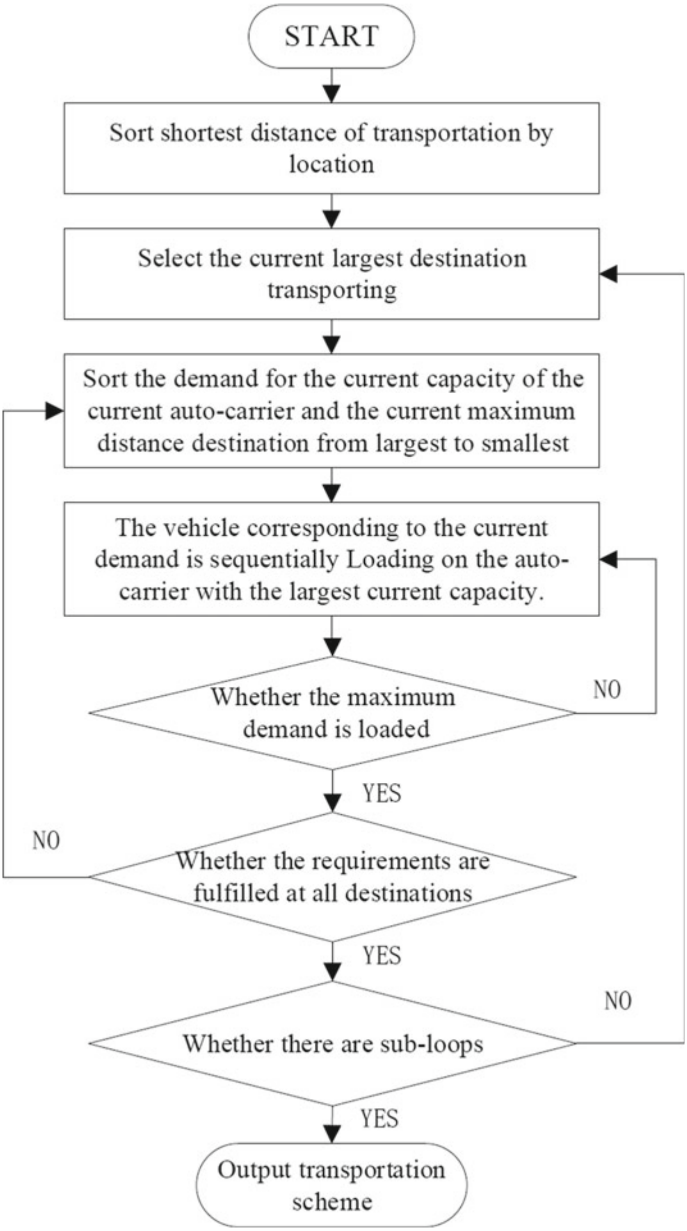

START
Sort shortest distance of transportation by location
Select the current largest destination transporting
Sort the demand for the current capacity of the current auto-carrier and the current maximum distance destination from largest to smallest
The vehicle corresponding to the current demand is sequentially Loading on the auto-carrier with the largest current capacity.
Whether the maximum demand is loaded
NO
YES
NO
Whether the requirements are fulfilled at all destinations
YES
NO
Whether there are sub-loops
YES
Output transportation scheme

Fig. 3 The improved algorithm flow table

(3) *Set constraints*

(a) *Passing a certain point*

```
for i in xindex.keys():
    if i[0] != 0:
        model.addConstr(y.sum(i[0],i[2],'*') >= x[i])
```

b) Preprocessing and sub-loop constraints

```
    model._acnum = len(Auto_carriers)
    model._vars = x
    model.Params.lazyConstraints = 1
    model.Params.TimeLimit = 100
    model.optimize(subtourelim)
    for i in xindex.keys():
       if x[i].x > 0.9:
        if Auto_carriersSort[i[2]-1] not in Routes.keys():
           Routes[Auto_carriersSort[i[2]-
 1]]=[(i[0],i[1])]
        else:
     Routes[Auto_carriersSort[i[2]-1]].append((i[0],i[1]))
```

In the specific solution process, pre-processing is used to achieve the height constraint (7), and callback is used to solve sub-loop constraint (8). Aiming at the above multi-objective combination optimization problem, we adopt the delaminating sequence method to list the above-mentioned objective functions according to their importance. Then find the optimal solution for the next target in the optimal solution set of the previous target each time until the common optimal solution is obtained.

5 Example Analysis

In order to testify the effectiveness of the method, randomly selected ten points from a regional transportation company. In In the process of transportation, the collecting vehicle must not exceed the maximum loading capacity of the auto-carrier, otherwise it will be suspended. Therefore, the company's transportation vehicles loading problem is a closed vehicle routing problem based on the loading capacity constraint. According to its current route and the models and algorithms proposed above, the current vehicle route was re-optimized.

The implementation process is mainly given the heterogeneous dimensions of the auto-carrier based on the central warehouse, and the set of vehicles required by each dealer. Then the vehicles are loaded into the auto-carriers and transported along the road network. The usage cost of auto-carrier is determined by their types and total kilometers traveled. The goal is to minimize costs, travel distances, and usage number of auto-carriers. The proposed procedure is composed by four main steps that are performed in cascade.

Step 1: Enter location data in the solver, including the customer's name and the latitude and longitude of the location. The data mainly indicate the starting and ending positions of the auto-carrier and the delivery point of the vehicles. See Table 1 for parameters.

Step 2: Enter the auto-carrier data, which mainly includes data such as auto-carrier name, model, starting and ending points of the transportation process, the cost of using the auto-carrier, and the length and width of the upper and lower floors respectively. See Table 2 for parameters.

Step 3: Enter vehicles data, which mainly includes data such as vehicle name, delivery locations for different customers, vehicles length, width, and height. See Table 3 for parameters.

Step 4: According to the different sizes of vehicles, the most suitable auto-carrier is selected for distribution, and the loading plan and routing strategy and be obtained by the algorithm. Case test results have selected auto-carrier V1 and V5 for delivery.

Table 1 The location data

Location ID	Latitude	Longitude
L1	31.2319426	121.5416838
L2	31.2335620	121.4599379
L3	31.2218637	121.4534166
L4	31.2823439	121.4148036
L5	31.0713752	121.3999500
L6	31.2595938	121.4414861
L7	31.2450446	121.4780849
L8	31.2443530	121.4222697
L9	31.3128295	121.4664003
L10	31.2924484	121.4311156

Table 2 Auto-carriers data

Auto-carrier ID	Type	Start point	End point	Cost of use	Lower layer length (m)	Lower layer width (m)	Upper layer length (m)	Upper layer width (m)
V1	I	L1	L1	10	19	2.7	19	2.7
V2	II	L1	L1	18	25	2.7	25	3.5
V3	III	L1	L1	30	19	3.5	19	3.5
V4	II	L1	L1	18	25	2.7	25	3.5
V5	II	L1	L1	18	25	2.7	25	3.5

Table 3 Vehicles data

Vehicle ID	Location point	Length (m)	Width (m)	Height (m)
C1	L7	4.60	1.75	1.50
C2	L8	3.99	1.64	1.68
C3	L2	4.53	1.77	1.70
C4	L4	4.43	1.64	1.68
C5	L3	3.76	1.79	1.80
C6	L7	4.78	1.68	1.63
C7	L3	4.06	1.60	1.42
C8	L2	4.77	1.73	1.52
C9	L10	4.54	1.63	1.76
C10	L4	4.47	1.80	1.60

Table 4 Delivery plan of auto-Carrier V1

Auto-carrier ID	Location ID	Vehicle ID
V1	L1	–
V1	L8	C2
V1	L4	C10
V1	L4	C4
V1	L10	C9
V1	L2	C3
V1	L3	C5
V1	L3	C7
V1	L2	C8
V1	L1	–

The delivery plan of auto-carrier V1 is shown in Table 4, and the delivery plan of the auto-carrier V5 is shown in Table 5.

The actual data confirms the correctness of the model, that is, accurate optimization values can be obtained by solving the solver. In addition, according to the survey of the company's situation and the manager's interview, it can be seen that the past

Table 5 Delivery plan of auto-carrier V5

Auto-carrier ID	Location ID	Vehicle ID
V5	L1	–
V5	L7	C1
V5	L7	C6
V5	L1	–

business volume usually requires more than 3 auto-carriers to be fully transported. We have reduced the number of auto-carriers that have completed the task, and because of the planned path, the auto-carrier V5 has a short travel distance and can be devoted to the next transportation task in time. And V1 fully meets the needs of customers at multiple points in one task. It is proved that our optimization model and precise algorithm can effectively solve the bottleneck problem of 3PL enterprise vehicle transportation.

In order to obtain the optimal combination of multiple calibration schemes, a test program for specific data is also written, so that as many simulations as possible for the placement of each layer of the optimal combination scheme of the auto-carrier can be obtained, and more practical schemes can be implemented. If the amount of the three types of auto-carrier meets the transportation requirements, and the number of auto-carrier consumed meets this set of data in this simulation sequence, it means that this set of data meets the requirements. Then output this set of data and its detailed arrangement scheme. After that, continue to do the rest of the simulation with this set of data, and output other combination schemes under this set of data. After the simulation of this set of data is completed, the number of three types of coupe models for the next set of simulations is automatically cycled. Finally, we got the type1-1 auto-carrier V1 and type1-2 auto-carrier V5 solution that meet the requirements.

6 Conclusion

The length, height, width, safe distance, and other loading restrictions of the auto-carrier as important factors that affect its space utilization and transportation efficiency. Based on the literature research, we found that although many literature have obtained fruitful research results in vehicle routing optimization. However, few literature puts attention on the optimization of vehicle logistics path planning with loading constraints, especially in the case of multiple orders [6]. It is necessary to consider not only the loading optimization and distribution route of a single auto-carrier and a single vehicle, but also the combination optimization and route selection of multiple auto-carriers and multiple vehicles. There are few studies on how to consider the loading limit of the auto-carrier, how to improve the loading rate of different car transporters in batch orders, and how to eliminate the sub loop in one-way transportation of different orders [7].

Hence, we construct a mixed integer programming model that based on the idea of integer programming including multi-constraints and multi-objectives, such as loading constraints, path balance constraints, and sub-loop constraints. The mathematical programming solver Gurobi and some heuristic methods are applied. The shortest path is selected and calculated using Floyd, and multi-objective optimization is solved by the layered weight method. It solves the combination of loading decision, vehicle selection and route optimization during the transportation process of the auto-carriers, and experiments are performed based on the historical order data of an enterprise. The experimental results show that this scheme can realize the scheduling of reducing the use of auto-carriers and achieving the scheduling optimization goal of logistics transportation cost. It provides a reference for vehicle logistics companies to optimize the path of auto-carriers transportation problem under loading constraints.

Acknowledgements This work was partially supported by Beijing Logistics Informatics Research Base.

References

1. H.-K. Chen, A heuristic algorithm for the auto-carrier loading problem. Int. J. Shipp. Trans. Logist. **8**(1), 21–50 (2016)
2. G.Y. Agbegha, R.H. Ballou, K. Mathur, Optimizing auto-carrier loading. Transp. Sci. **32**(2), 174–188 (1998)
3. R. Tadei, G. Perboli, F. Della Croce, A heuristic algorithm for the auto-carrier transportation problem. Transp. Sci. **36**, 55–62 (2002)
4. A.C. Bonassa, C.B. Cunha, C.A. Isler, An exact formulation for the multi-period auto-carrier loading and transportation problem in Brazil. Comput. Ind. Eng. **129**, 144–155 (2019)
5. J.-F. Cordeau, M. Dell'Amico, S. Falavigna, M. Iori, A rolling horizon algorithm for auto-carrier transportation. Trans. Res. Part B, **76** (2015)
6. H. Wang, Q. Meng, X. Zhang, Multiple equilibrium behaviors of auto travellers and a freight carrier under the cordon-based large-truck restriction regulation. Transp. Res. part E: Logist. Transp. Rev. **134**, 101829 (2020)
7. IORI, Manuel, An annotated bibliography of combined routing and loading problems. Yugoslav J. Oper. Res. **23**(3) 2016

Towards Using Micro-services for Transportation Management Systems

Sabah Mohammed, Jinan Fiaidhi, and Mincong Tang

Abstract The growth of today's cities and the increased population mobility are providing great challenge to manage vehicles on the roads. This challenge led to the need for new and innovative traffic management, including the mitigation of road congestion, accidents, and air pollution as well as many business oriented demands. Over the last decade, researchers have been focusing their efforts on leveraging the recent advances in Web Services and Multi-Agents to design new road traffic management systems (TMS) for resolving these important challenges in the future transportation. However, these new solutions are still be insufficient and complex to construct TMS systems that are capable of handling the anticipated influx of the population, vehicles and changing transportation scenarios. This paper is pointing to a new and emerging technology that can solve these challenges and develop more flexible TMS systems based on the notion of microservices offered by web frameworks like IFTTT, Zapier, Node-Red and WoTKit.

Keywords TMS · Web services · Microservices · Node-red · IFTTT · Zapier · Node-Red · WoTKit

1 Thinking Out of the Box

Transportation management systems (TMS) are one of the core technologies used in supply chain management (SCM). Traditionally, these systems are available as standalone software or as modules within enterprise resource planning (ERP) suites. TMS serves both shippers, carriers and other logistics providers including distributors,

S. Mohammed (✉) · J. Fiaidhi
Department of Computer Science, Lakehead University, Thunder Bay, Canada
e-mail: sabah.mohammed@lakeheadu.ca

J. Fiaidhi
e-mail: jfiaidhi@lakeheadu.ca

M. Tang
School of Economic and Management, Beijing Jiaotong University, Bejing, China
e-mail: mincong@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_8

Table 1 Popular TMS systems

TMS system	Website
3Gtms	https://www.3gtms.com/
CEVA Logistics	https://www.cevalogistics.com/
BluJay	https://www.blujaysolutions.com/
Cloud Logistics	https://www.gocloudlogistics.com/
Descartes	https://www.descartes.com/
JDA	https://jda.com/
Kuebix	https://www.kuebix.com/
Manhattan	https://www.manh.com/
MercuryGate	https://www.manh.com/
Oracle	https://www.oracle.com
SAP	https://www.sap.com
TMC	https://www.mytmc.com/
TMW	https://www.tmwsystems.com/
Transplace	https://www.transplace.com/

wholesalers and retailers. TMS systems have gained traction over the past decade as an enabler of global trade and logistics. Gartner, in its March 2019 Magic Quadrant report,[1] predicted the global TMS market will grow at an accelerated rate, reaching $1.94 billion by 2022. The use of TMS systems reduce transportation complexity dramatically by integrating fleet and logistics management throughout the transportation network. Table 1 lists some of the notable TMS systems currently in use.

In principle TMS can benefit companies in numerous ways, including the following:

- Improved Cash Flow
- Minimal Paperwork
- Inventory Management
- Supply Chain Visibility
- Optimized Routing through Pool Distribution
- Tracking Drivers En-Route
- Accurate Order Fulfillment
- Improving Customer Experience
- Reduced distribution and warehouse
- Reduced administrative costs and invoicing errors.

However, TMS ecosystem is changing and there is a timely need to incorporate further advancements and updates to the current systems. TMS needs to adapt to the

[1] https://www.gartner.com/en/documents/3905867/magic-quadrant-for-transportation-management-systems.

new ecosystem changes to ensure that a shipper is able to succeed in the next decade. Among the new trends are the followings:

- Adoption of cloud-based system,
- More load optimization automation
- Adoption of open-sourced platforms
- Enforcement of connected truck using more Internet of Things technologies
- Adoption of blockchain-Driven Software to encourage partnership
- Enhancement of Real-Time Supply Chain Management for greater visibility
- Intuitive Management and Analytics capabilities via integrating more Machine Learning and AI components
- Enhanced Web Portals.

These trends present huge challenges that shippers face in today's hectic freight transportation environment which advocate for more integration, flexibility and high responsiveness. The present remedy to outsource the added value components that cannot be captured by the enterprise TMS to third party venders and use sound third party logistics (3PL)[2] software to manage dealing with the third party partners. The 3PL software is required to have the ability to integrate with the major e-commerce platforms as well as satisfying the following constraints:

- Manage contracts and service level agreements for SCM outsourcing
- Provide features for multi-warehouse and inventory management
- Deliver functionality for transportation management and shipping
- Track the costs (estimate and actual) of all outsourced supply chain activities
- Maintain a repository of providers of supply chain operations
- Allow customers to access information such as inventory availability
- Include metrics that customers can use to track performance.

However, selecting a 3PL software to compliment or integrate with the shipper TMS requires careful thought and great expertise. Integrating the 3PL with TMS/ERP systems can bring enormous efficiencies. If these integrations are out-of-the-box, they can be implemented quickly so that we can see the return of investment (ROI) immediately. By integrating the 3PL and TMS/ERP, we'll reduce the need to re-enter information between systems and facilitate the rapid creation of shipments with correct documentation.

Moreover, this integration should also be able to scale with the business when it needs change. All the above constraints increase the complexity of having modern transportation ecosystem that is inherently a network-based business process. It involves an ecosystem of different parties—a community of shippers, carriers, consignees, brokers, and others that need to communicate and collaborate with each other in order to transport products and utilize assets and labor as efficiently as possible. This complexity and fragmentation makes it challenging to quickly and efficiently match transportation demand with available vender capacity of whatever size. This growing need in the market for better matching of supply and demand,

[2] https://www.g2.com/categories/3pl.

coupled with the rise of cloud computing, web services, software intelligent agents, software-as-a-service (SaaS), modern application programming interfaces (APIs), and other emerging technologies including microservices, is driving the next evolution of transportation management systems [1]. This ample demand for technology in supply chain management and rapidly growing supply of TMS/3PL tools aiming to improve the industry, have failed to connect the two in a way which truly adds value. This paper aims to guide logistics professionals open to innovation driven by technology but eager to avoid common pitfalls.

2 Innovation in the Transport Logistics

Innovation has historically played a vital role in increasing efficiency. However, while other industry sectors have experienced rapid growth of productivity, the transport logistics industry has seen relatively small improvement in terms of efficiency [2]. Increasing transport logistics efficiency through identifying the right innovation technologies can save an enormous amount of money, hence is worthy of study. Notable attempts based on important innovations were based on *web services* [3] to develop a freight brokering system. Web services are self-contained, Web-enabled applications capable not only of performing business activities on their own, but also possessing the ability to engage other Web services in order to complete higher-order business transactions. Developing freight brokering system is an attempt to match the logistic capabilities of transportation providers with the needs of merchandise owners. From a service science perspective, the clients and transporters interacting through the matchmaking system create mutual value by minimizing client costs, maximizing business opportunities, increasing competition, and reducing operational risks. Shifting the dominant thinking of supply chain management toward the concepts of web services ecosystems opens up many research opportunities and strategies for improved organizational performance [4]. However, the key to solving the problem of efficient management of logistics information is centered on building a smart logistics services supply system. In this direction several researchers proposed a *multi-agent system* for smart brokering of logistics web services (e.g. allowing shippers to collaborate during the logistics process, thus improving the overall system intelligence; the use of agent-based negotiation for logistics management). As an example, in [5] the authors have proposed composing web services using coalitions of agents performing tasks for service requestors.

Although the use of web services and multi-agent technologies in designing smart brokering logistics is motivated by the interactivity, interoperability, responsiveness and social characteristics of this domain, these technologies uses complex development frameworks that uses an Extensible Markup Language (XML)-based service description for the description of each web service. Such a description covers all the necessary details needed in order to interact with the services, including message formats, transport protocols, and methods. Web services share Universal Description and Ontologies, Discovery and Integration (UDDI), which is a centralized service directory, for its service discovery. Multi-agents technology on the other hand provides capabilities of autonomy, social ability, reactivity, and the ability to be proactive. Developing a multi-agent system requires that each functional unit (e.g., the firm) to be populated by a number of heterogeneous intelligent agents with diverse goals and capabilities. Each agent, then, is designed to represent a specific functional unit. The requirements for the action strategies and policies to be used may be entered into the agent beforehand. Different levels of collaborative requirement can be easily incorporated into the agent as different goals are made based on different types of scenario. Multi-agents are very effective in addressing both coordination and conflicts among the firms. Thus multi-agents require to have demand responsive components that provide a share transportation services with flexible routes and focus on optimizing of economic values. The design of such demand responsive components is still in its early stage to become part of the smart TMS systems [6]. Figure 1 illustrates the innovative technologies that have been suggested so far for designing a smart transportation system.

But the picture is changing rapidly related to the emerging innovations in transport technologies. Recently, new opportunities for developing lightweight transport brokering systems are arising, as Web of Things (WoT) platforms such as IFTTT [7], Zapier [8] and AWS Lambda Architecture [9] are emerging. These platforms

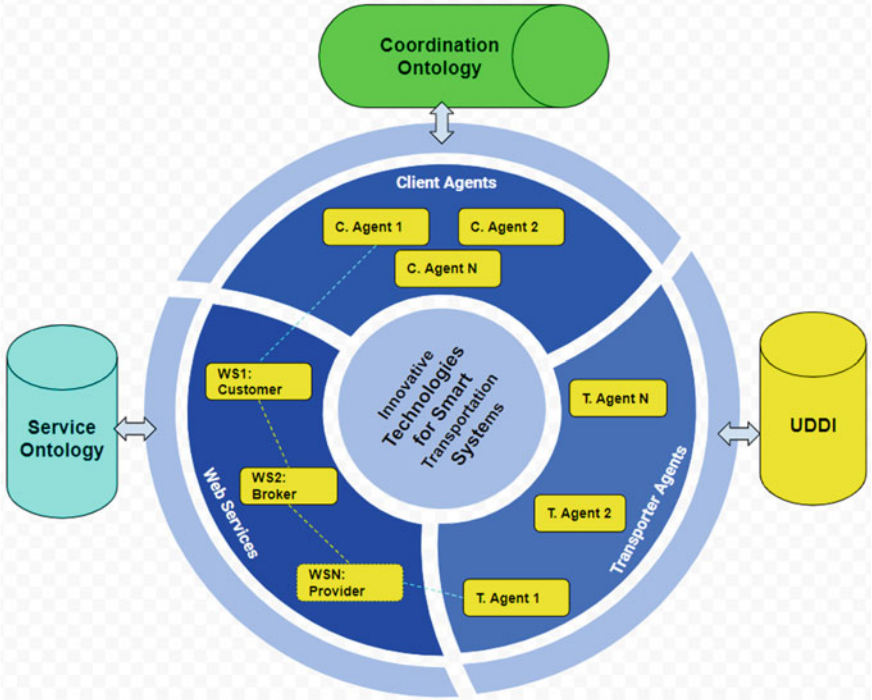


Fig. 1 Innovative technologies for designing logistic transport brokering system

came into service to support flexible composition of applications with various things connected to the Web and over the cloud. A shipper can easily select an application component from a pool of building blocks such as sensor information, actuation functions and data services to create and deploy personalized transport applications. We can reasonably expect more Web applications to be created through such WoT platforms because of the ease of development of large applications in the cloud as a set of small services that can be independently developed, tested, deployed, scaled, operated, and upgraded. These emerging technologies are based on using microservices [10]. Microservices architecture involves breaking down a software application into its smaller components, rather than just having one large software application. Typically, this involves splitting up a software application into smaller workflow components. These workflows can then integrate to each other via an interface. Microservices also present a solution to scaling large applications that was the problem associated with most of the legacy monolithic web applications. Monolithic applications need to be designed to accommodate any increase in the demand even if the demand is for one highly consumed service. Therefore, given a large infrastructure to accommodate anticipated demand that is shared among services, server resources will provide a wasted in the execution of unused services, thus increasing related costs. To avoid the problems of monolithic applications and take advantage of some of the service oriented (SOA) architecture benefits, the microservice architecture pattern has emerged as a lightweight subset of the SOA architecture pattern [9].

3 Towards Microservice Based Transport Systems

The execution pattern of a microservice application can be described as a time sequence of Web service invocations. Each web service is categorized into either a trigger or an action in WoT architecture [11]. A trigger is either a publication of some information or a signal that an action (actuation) took place. An action is a task to be executed whenever a trigger is fired. In this approach, instead of producing Apps as a complete runtime artifact, developers can design intermediary artifacts, as accessible controls parts for typical transportation tasks. Using these controls, transporters are enabled to construct their own tasks autonomously by using resources of their own choice from across the WoT. The microservice based transportation system translates the time sequence of trigger and action executions to a time sequence of network flows. A network flow is a traffic information between two communicating endpoints. Zapier and IFTTT or many other alternatives[3] makes this task easy. It lets transporters to integrate everyday apps (e.g. existing legacy app or a newly created workflow) and automate the transportation business processes. Figure 2 illustrates the idea behind using WoT platforms.

[3] https://www.gadjetgeek.com/zapier-alternatives-open-source/.

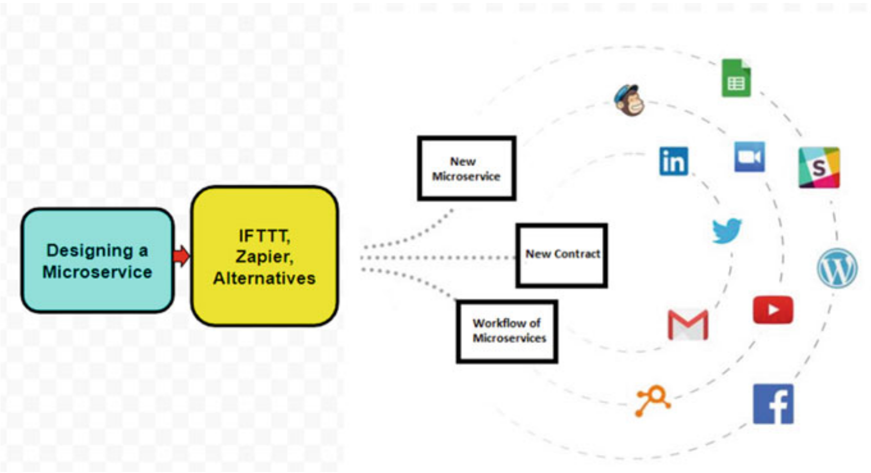
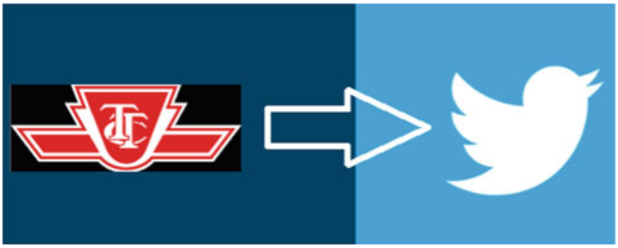


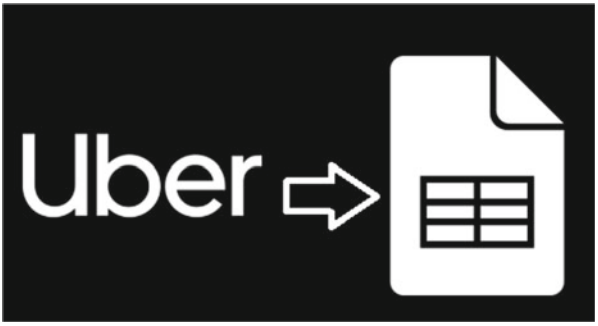
Fig. 2 Developing microservices based on WoT platforms

WoT platforms constructs a new microservice by combining Triggers and Actions via completing an action in one app when a trigger occurs in another app. These combos—called "Zaps" according to Zapier terminology or "Applets or Recipes" according to IFTTT terminology— will complete the workflow tasks automatically. Using the IFTTT platform, the "IF THIS" keyword is used to identify the triggers and "THEN THAT" is used to identify the actions part. The trigger is activated by changes that occur within other web services such as TTC[4] (Toronto Transit Alert). Upon activation you may assign an action like pushing a notification about the transportation alert to your twitter account:

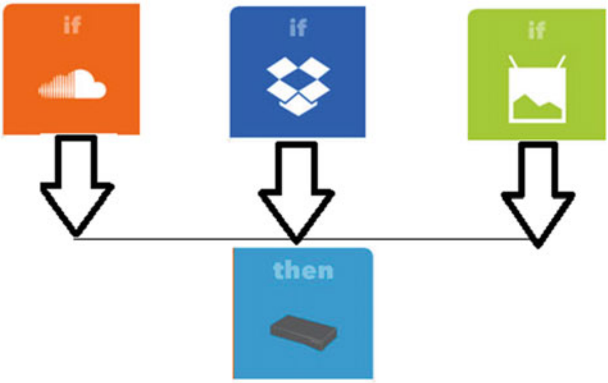


WoT platform may create a new transportation contract upon firing a new microservice. For example log your completed Uber trips in a spreadsheet:

[4] https://www.ttc.ca/Service_Advisories/all_service_alerts.jsp.



Moreover, we can add branching logic to microservices to run different actions based on different conditions (e.g. if A happens in your trigger app, then do X. If B happens, then do Y, and so on). Tools like Paths[5] can be used with Zapier to add such branching logic and produce a workflow of microservices (e.g. Case downloads are songs, normal files or Android images Then save to the Seagate Personal Cloud):



WoT platforms like IFTTT or Zapier, however, may represent decent workflow composition tool to get started on simple point-to-point integrations but what should you do when you want to grow your organization and tackle the complexity that often arises when doing so? Fast-growing transportation companies and enterprises need to incorporate higher level frameworks to accommodate the various complex scenarios. In this direction we can use more sophisticated workflow wrapper like the WoTKit[6] and Node-RED.[7] With these wrappers one can produce fully customizable workflows with flexible connector operators such as loops, data storage, array mapping, branching, and if/then conditionals and many more. Node-RED is a programming tool used for wiring together hardware devices, APIs and online services by representing them as nodes. It has been recently used in designing a smart TMS system

[5] https://zapier.com/paths.

[6] https://wotkit.readthedocs.io/en/latest/.

[7] https://nodered.org/.

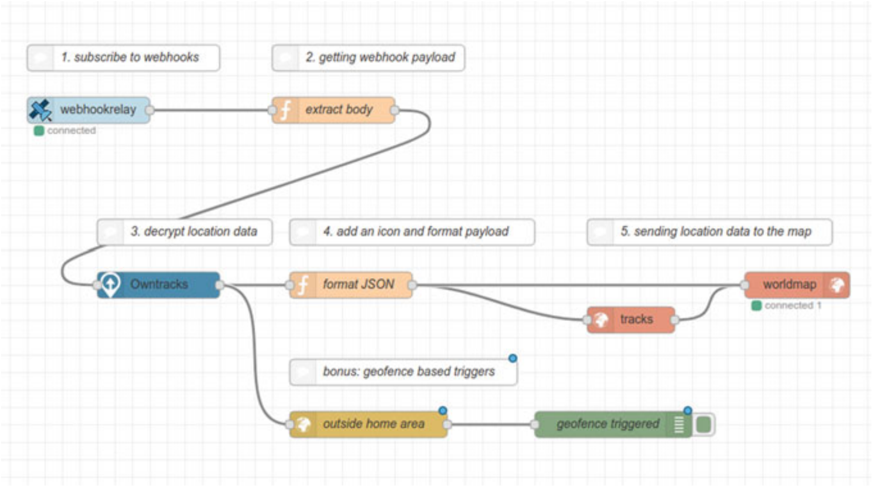


Fig. 3 Node-red microservices workflow for trucking drivers and vehicles

[12]. The advantages that render the uniqueness and innovation of Node-RED, as a programming tool, is its ability to amalgamate a widely-used programming language under a new flow-based programming model along with a large list of software and hardware components, retaining at the same time, an open-source character. In addition, the Node-RED UI offers a user-friendly way for creating and inserting new nodes with drag and drop—like methods. As a result, adding and connecting new components becomes trivial. It's obvious that an already implemented node can be reused in the same or another flow or even saved for future usage. Flows can be open-sourced or sold, enabling developers this way to focus on the innovative part of their task saving time on developing already implemented flows. For example if we want a Node-Red TMS system to have the capability to truck the location of their transportation vehicles, then we can transmit the driver mobile phone location via the OwnTracks[8] microservice to the Node-Red server. We can also visualize the vehicle position using worldmap[9] microservice. We can connect also ingest data from the OwnTracks microservice with geofence[10] microservice to create geofence based triggers. This added microservice is to create automated actions that get triggered when trackers leave or enter areas (for example the vehicle leaving the town while the driver phone is at home). Figure 3 illustrate the composite Node-Red workflow of this application [9]. The overall architecture of TMS systems based on the notion of microservices can be represented in Fig. 4.

[8] https://owntracks.org/.

[9] https://flows.nodered.org/node/node-red-contrib-web-worldmap.

[10] https://flows.nodered.org/node/node-red-node-geofence.

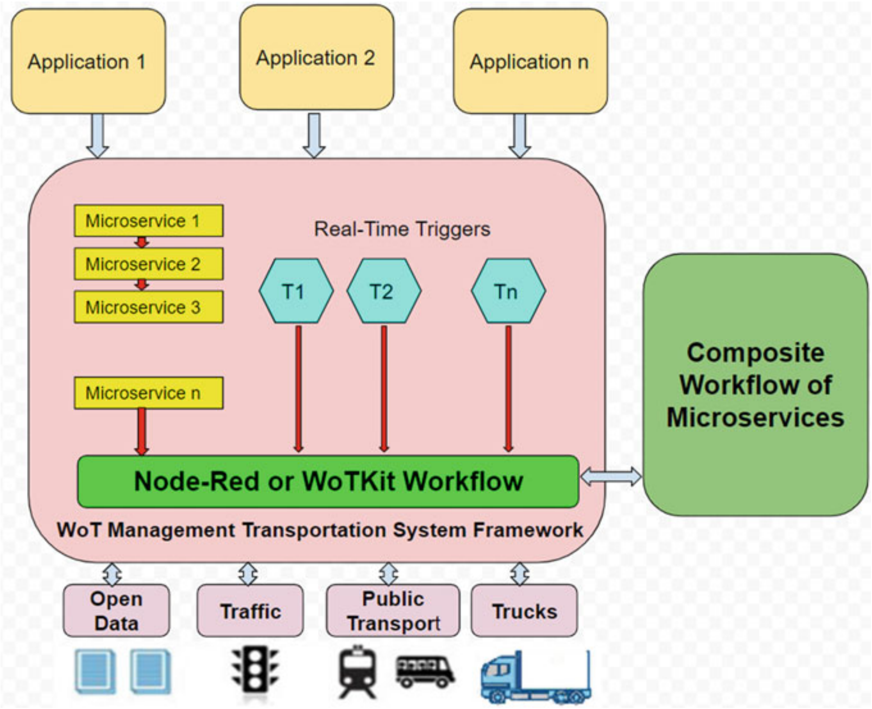


Fig. 4 The envisioned WoT MTS architecture

4 Conclusions

It is becoming more and more pervasive, modern web applications and platforms to consume or provide data and new functionalities as web services through APIs which you can access with simple calls, like an HTTP request. WoT frameworks, following this trend in much of its architecture, consists of autonomous components which are connected with each other by exposing their services in this way. So, instead of presenting raw data and other specialized services, a WoT platform is flexible enough to allow 3rd parties to develop complex transport microservice applications through the provision of frameworks like IFTTT, Zapier, Node-RED or WoTKit, can be a great asset for the modern MTS infrastructure. This paper illustrates this new trend in developing transportation management system based on the notion of microservices. This research is an ongoing research work in progress to develop innovative transportation system that are flexible enough to respond to the current demand and technology change.

Acknowledgements The authors would like to thank the IEEE LISS 2020 conference organizers (July 25–28, 2020 Budapest, Hungary) for the great work in making this event at this difficult COVID-19 outbreak.

References

1. I. Harris, Y. Wang, H. Wang, ICT in multimodal transport and technological trends: unleashing potential for the future. Int. J. Prod. Econ. **159**, 88–103 (2015)
2. D. Cipré, L. Polo, A. Capella. Innovation in transport logistics—best practices from the EU project LOGINN, in *Dynamics in Logistics* (Springer, Cham 2016), pp. 599–608
3. C. Ji,, M. Li, L. Li, Freight transportation system based on web service, in *IEEE international conference on services computing*, (SCC 2004). *Proceedings* (2004), pp. 567–570
4. R.F. Lusch, Reframing supply chain management: a service-dominant logic perspective. J. Suppl. Cha. Manage. **47**(1), 14–18 (2011)
5. L. Luncean, C. Bădică, A. Bădică, Agent-based system for brokering of logistics services–initial report, in *Asian Conference on Intelligent Information and Database Systems* (Springer, Cham, 2014), pp. 485–494
6. S. Satunin, E. Babkin. A multi-agent approach to Intelligent Transportation Systems modeling with combinatorial auctions. Exp. Syst. Appl. **41**(15), 6622–6633 (2014)
7. S. Ovadia, Automate the internet with "if this then that" (IFTTT). Behav. Soc Sci Lib. **33**(4), 208–211 (2014)
8. M. Finch. Using Zapier with Trello for electronic resources troubleshooting Workflow. The Code4Lib J. **26** (2014)
9. M. Villamizar, O. Garcés, L. Ochoa, H. Castro, L. Salamanca, M. Verano, R. Casallas et al., Cost comparison of running web applications in the cloud using monolithic, microservice, and AWS Lambda architectures. SOCA **11**(2), 233–247 (2017)
10. S. Newman, *Building Microservices: Designing Fine-grained Systems* (O'Reilly Media, Inc 2015)
11. Y. Yoon, H. Jung, H. Lee, Abnormal network flow detection based on application execution patterns from Web of Things (WoT) platforms. PLoS ONE **13**(1), e0191083 (2018)
12. S. Sicari, A. Rizzardi, A. Coen-Porisini, Smart transport and logistics: a Node-RED implementation. Internet Technol. Lett. **2**(2), e88 (2019)

An Analytic Approach for Workers' Fatigue Examination Using RFID-Enabled Production Data

Yishu Yang and Ray Y. Zhong

Abstract With the advantages of long-distance contactless identification and data storage capacity, the use of radio frequency identification (RFID) technology in the fields of manufacturing, transportation and logistics has been widely reported. Fatigue of workers plays a critical role in impacting the manufacturing efficiency because it reduces productivity and increases accident rates. Therefore, the workers' fatigue must be well examined and addressed. This paper thus proposes an analytic approach to use RFID captured production data and builds an effective method for mining the structural insight to predict the fatigue trajectory in workplace from a huge number of RFID data which may be full of inaccurate, incomplete and missing records. In this research, realistic processing time is used to measure the workers' fatigue. Based on a general framework for the fatigue examination, the proposed approach is able to estimate the employees' fatigue trajectory within designated period of time using RFID-enabled production data. Different genders and shifts are considered to find the key impact factors on fatigue.

Keywords Data-driven approach · Fatigue trajectory · Radio frequency identification (RFID)

1 Introduction

Larges numbers of occupational populations have been exposed to work-related fatigue risks in recent years [1]. It is of great significance to explore the evaluation and management of workers' fatigue to protect the physical and mental health of the occupational population. Fatigue generally refers to a state of physical insufficiency with specific conscious symptoms such as burnout and diminished energy as the

Y. Yang · R. Y. Zhong (✉)
Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China
e-mail: zhongzry@hku.hk

Y. Yang
e-mail: annyys@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_9

activity (or labor) continues [2]. In the workplace, it can be divided into physiological fatigue and psychological fatigue. Physiological fatigue refers to the phenomenon in which the individual's physiological function declines, while psychological fatigue refers to subjective experience of fatigue and tiredness due to continued consumption of cognitive and emotional will [3]. During manufacturing stage, fatigue is a multidimensional construct, and both phenomena mentioned above are included [4]. It is associated with many adverse consequences, including physical discomfort, decreased performance, productivity, quality of work, and increased workplace accidents and human errors [5]. Existing studies have shown that specific factors in work arrangements can promote fatigue [6]. The problem is how to assess workers' fatigue levels in a real-time basis and how to determine the key factors that cause fatigue.

Many laboratory-based studies have studied the effects of fatigue, and they simulate typical construction tasks such as elevated work, repetitive hammering, and nailing, sawing and screwing. However, very few research has focused on the pattern of fatigue development over time in work tasks [5]. Typically, when assessing the impact of workplace activities on fatigue levels, it uses qualitative self-reporting methods rather than quantitative data measurements. And it can prone to response bias while using this type of data collection in the workplace. Wearable technology has been proposed as a possible solution because it can collect objective physiological data such as steps and heart rates [7, 8]. However, the practicality of this device has yet to be proven and can be a problem. Since RFID technology has been widely used to collect a variety of manufacturing data to support shop floor production management, the real-time data can be used for fatigue analysis and prediction in RFID-enabled production environments.

This paper tries to offer a better opportunity to integrate the understanding of the psychology and physiology of the fatigue effectively and practically and provides some referential value for the fatigue risk management strategies. The main objective is to predict worker's fatigue trajectory during manufacturing using RFID-enabled shop floor production data and analyzed the curves obtained from different factors to support the fatigue management decision.

The reminder of this paper is organized as follows. In Sect. 2, some necessary backgrounds on related work are briefly introduced, including fatigue in the workplace and application of RFID technology in manufacturing. Then we present the approaches of data processing and the methods of fatigue examination in Sect. 3. In Sect. 4, we provide our results and discuss the results obtained from different factors to find the key influence. In the last section, we present the conclusions and future research directions.

2 Literature Review

2.1 *Fatigue in the Workplace*

Fatigue is a multi-system dysfunction that affects both physiology and cognition, and this has been widely studied in observational experiments in the workplace and laboratory environment [9]. The impact of fatigue on cognition, including inattention, uncoordinated movements, unresponsive, memory impairment and more frequent loss of concentration [10]. Existing studies have shown that specific factors in work arrangements can promote fatigue [6]. Therefore, Fatigue can be managed like other risk factors. To control the fatigue and reduce the risk of accidents, some safety-critical industries have strict supervision of workers: airline staff, maritime crew, road transport personnel and operators of public service vehicles must adhere to these rules [11].

Fatigue can be induced by a variety of interrelated factors. The lack of sleep and interruption of the normal sleep cycle are considered to be some of the critical factors in causing fatigue [6]. In the workplace, fatigue is caused by the workload of employees, and it affects employees' fatigue levels from three aspects: physical, environmental and spiritual [12]. Sadeghniiat-Haghighi and Yazdi [1] gave some measurements and proposed a new and more comprehensive method called fatigue risk management system (FRMS) which is based on sleep science. It considers the effects of sleep on fatigue levels and improves worker fatigue through sleep-related management measures. Two possible approaches (fatigue reduction and fatigue proofing) are used to control fatigue, and they are implemented as an auxiliary implementation of the FRMS to better manage fatigue [1]. However, as mentioned above, the influencing factors of fatigue are various and complex. It is affected by multiple factors. In addition to sleep, other influencing factors should also be considered, such as gender, age, and work shifts. FRMS is a general control of fatigue based on sleep factors. It does not consider the impact of different working hours and other factors on employees. To achieve better control of employees' fatigue, the methods should be targeted.

Greig and Snow [11] has studied fatigue in the medical care industry. The article argues that the working hours of medical staff should be reduced, and a shorter shift system should be implemented to reduce the fatigue of medical staff. It encourages doctors to determine when they are most vulnerable and need the rest through the warning signs of the railway industry that need to review the work mode and risk management in [10]. Reduction of working hours can reduce fatigue, but whether it is in the medical care industry or the manufacturing industry, shorter shift systems may require more staff, and the subsequent cost issues need to be considered. In addition, relying on employees to make their own fatigue judgment may not be an accurate method, and will include many subjective estimations. Employers in the manufacturing industry may not accept with this fatigue assessment method. By training the employees who arrange the shift schedules to obtain a more scientific staff shift schedule, thereby improving fatigue management [11]. This is a more desirable method, but there is still doubt about what kind of training employees should conduct, and what is a scientific schedule is difficult to judge.

2.2 Application of RFID Technology in Manufacturing

RFID technology enables non-contact, two-way communication through induction, radio waves or microwave energy to achieve the purpose of identification and data exchange. This technology can perform long-distance recognition on high-speed moving objects, and is a non-contact automatic recognition technology [13]. With the advantages of long-distance contact-less identification and data storage capacity, RFID technology is widely used in manufacturing, transportation, logistics and other fields. How to obtain information in the process of manufacturing informatization and how to effectively manage the obtained information in a complex production environment accompanied by changeable production data and interlaced related information are the key problems faced by many enterprises. It is also a crucial step in solving information services [14]. The intelligent data acquisition terminal developed with RFID technology can meet the requirements of the field of machinery manufacturing such as data display, data input, and data transmission [15].

Article [16] analyzed three types of fresh agricultural products (FAP) supply chain models. By establishing a revenue model before and after using RFID technology, it is concluded that the application of radio frequency identification technology can reduce the loss rate of FAP transportation, increase the production cost and the recycling rate of RFID tags. However, this article only considers the impact of RFID technology on transportation time and loss rate. More factors need to be further studied. Reference [17] considered the value of RFID technology in the apparel manufacturing industry. According to [18], a textile company called Novelex uses RFID-based systems. In the system, 300,000 tags are attached to the products, and four sets of RFID readers are placed. This measure improves the accuracy of the data and also the delivery efficiency, reducing the total production preparation time from 35 days to 30 days. In reference [19], a Hong Kong based clothing manufacturer improved its clothing manufacturing through RFID technology, saving 5.1–10.3% of the order planning time. Reference [17] conducted further calculation experiments to determine the average percentage change in profits of the entire supply chain after the establishment of RFID technology. It is found that by reducing production lead times, RFID technology can significantly increase supply chain profits. This means increased production capacity throughout the supply chain and the ability to accept more orders from retailers.

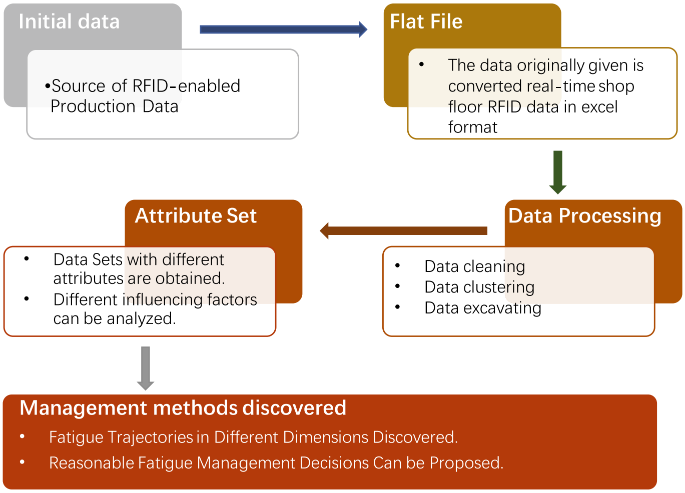
3 Methods

A general framework for the fatigue examination using RFID-enabled production data is shown in Fig. 1.

3.1 Data Collection

The data used in this paper is real data from a partner company that has used RFID technology to support its workshop production for 7 years. The real-time shop floor production data obtained by RFID technology includes complex manufacturing logic, which makes data pre-processing necessary. The obtained real-time data needs to be converted into understandable flat data before further data processing and analysis.

To avoid the impact of the original test and the frequent improvement of RFID technology on the acquired data, the earlier data is discarded. In this paper, the samples we selected for the experiment came from production data obtained in 2008. Here, the data originally given is converted real-time shop floor RFID data in excel format. The flat file consists of transactions for production operations and activities. Each transaction includes: ID, BatchMain ID, UserID, ProcCode, ProcSeqnum, Quantity, GoodNum, Time, Location. The definition of these attributes is shown in Table 1.



Initial data
- Source of RFID-enabled Production Data

Flat File
- The data originally given is converted real-time shop floor RFID data in excel format

Data Processing
- Data cleaning
- Data clustering
- Data excavating

Attribute Set
- Data Sets with different attributes are obtained.
- Different influencing factors can be analyzed.

Management methods discovered
- Fatigue Trajectories in Different Dimensions Discovered.
- Reasonable Fatigue Management Decisions Can be Proposed.

Fig. 1 A framework for the fatigue examination

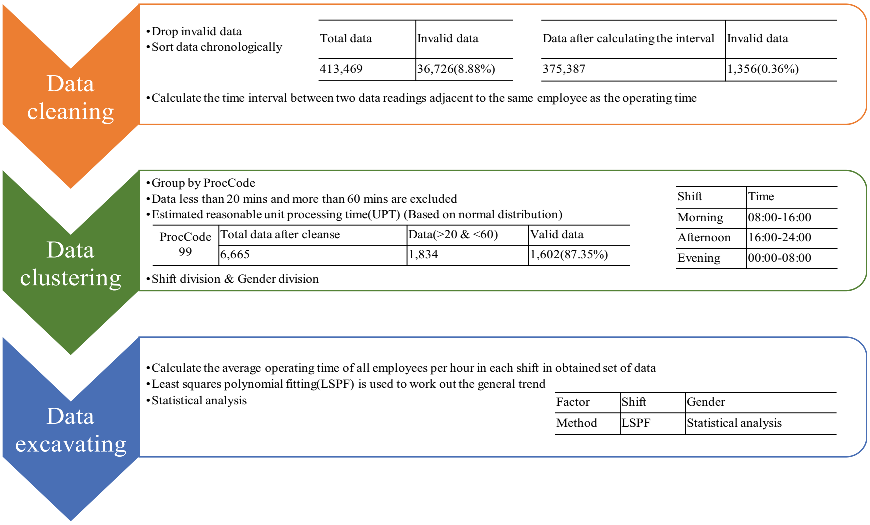
Table 1 Definition of the data code

Data code	Definition
ID	Product code
BathMainID	Required group of materials to manufacture a product
UserID	Individual worker/operator at a manufacturing plant
ProcCode	Type of mechanical manufacturing process
ProcSeqnum	Order in which each process is taken place
Quantity	Number of products produced
GoodNum	Quantity of product that passes the acceptance threshold
Time	Period of record at each event occurrence
Location	Representing a specific machine

3.2 Data Processing

The overall vision of data processing is shown in Fig. 2. Three steps are included in this process and each step is matched with the corresponding logic algorithm. The initial data size exceeds 23 MB, all of which were extracted from 2008 totaling about 410,000 pieces.

1. *Data cleaning*. The application of RFID technology in manufacturing has created a ubiquitous production environment, where fatigue analysis and prediction can



- Drop invalid data
- Sort data chronologically

Total data	Invalid data
413,469	36,726(8.88%)

Data after calculating the interval	Invalid data
375,387	1,356(0.36%)

- Calculate the time interval between two data readings adjacent to the same employee as the operating time

- Group by ProcCode
- Data less than 20 mins and more than 60 mins are excluded
- Estimated reasonable unit processing time(UPT) (Based on normal distribution)

ProcCode	Total data after cleanse	Data(>20 & <60)	Valid data
99	6,665	1,834	1,602(87.35%)

- Shift division & Gender division

Shift	Time
Morning	08:00-16:00
Afternoon	16:00-24:00
Evening	00:00-08:00

- Calculate the average operating time of all employees per hour in each shift in obtained set of data
- Least squares polynomial fitting(LSPF) is used to work out the general trend
- Statistical analysis

Factor	Shift	Gender
Method	LSPF	Statistical analysis

Fig. 2 Overall vision of data processing

be enabled. However, the data captured by RFID is huge and full of inaccurate, incomplete data as well as lost records [20]. In the first step, the blank data is eliminated. About 36,726 invalid data were discarded, accounting for 8.88% of the total. Secondly, all data are sorted in chronological order. The third step is to calculate the time interval between two data readings adjacent to the same employee and take it as the operating time (that is, the end time of the previous batch of production is used as the start time of the next batch of production). The feature set can be expressed as $DF = \{ID, BID, UID, PC, PS, Q, FT, MID\}$, A piece of that $df_i = \{id_i, bid_i, uid_i, pc_i, ps_{i,}q_i, ft_i, mid_i\}$, and the actual operating time of a product batch could be calculated through $ft_{i+1} - ft_i$. After calculating the interval, 375,387 data remained. Finally, after the calculation, eliminate all invalid data.

2. *Data clustering*. The cleaned data needs to be clustered with different features so that the influence of different factors on the research object can be found.

   (a) *Group by*. The cleaned data is grouped according to the type of mechanical manufacturing process (The standard operating time required by different manufacturing processes is greatly different. For example, completing the fine grinding process of a product often takes several tens of times of the rough grinding process).

   (b) *Mining operating time*. Take one set of data and the normal distribution of operating time is shown in Fig. 3. For fine grinding, the average operating time for 10 employees in the taken set of data has been calculated before and the data less than 20 min and more than 60 min can be excluded as invalid data. In this paper, three sets of data (ProcCode99, 118, 97) were taken out for analysis and comparison. ProcCode99 is used as an example in this chapter. It can be seen from Fig. 3 that more than 80% of the data are

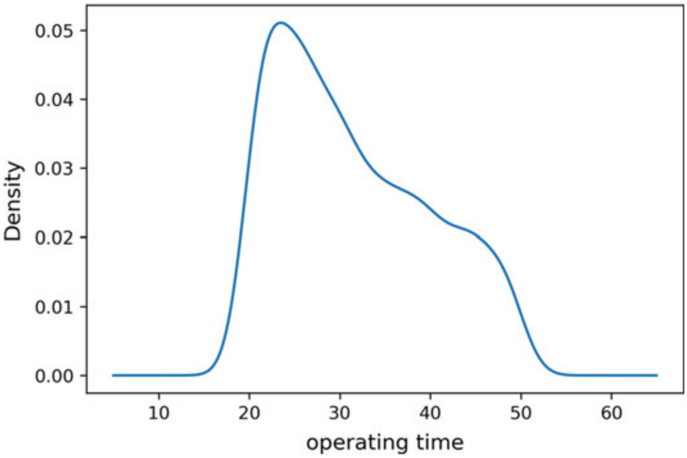


Fig. 3 The normal distribution of operating time (ProcCode99)

in the range of 20–50. Based on this, data between 20 and 50 are considered valid data.

(c) *Shift division.* Three shifts system is adopted with 8 h per shift. That is, the data is divided into three shifts: morning, afternoon, and evening. To simply process the data and discover regular patterns, in this paper, the morning shift is from 8:00 to 16:00, the afternoon shift is from 16:00 to 24:00, and the evening shift is from 00:00 to 08:00. Three sets of data were divided into shifts for analysis and comparison, which were rough face grinding (ProcCode97), fine face grinding (ProcCode99), and fine tapered face grinding (ProcCode118).

(d) *Gender division.* Gender clustering was performed on the ProcCode99 data set. Data from 26 employees were selected as the gender factor research data set, including 13 men and 13 women.

3. *Data excavating.* In order to dig out the influence of different factors on the research object from the classified data set, further calculation and fitting of the data set are needed. Least squares polynomial fitting is used to work out the general trend. It can be mathematically proved that any function can be expressed in polynomial form. Statistical analysis is also used.

4 Data Analyses and Discussion

The unit processing time of the product is positively related to the employee's fatigue level. That is, the higher the employee's fatigue level, the longer the unit processing time. Here, we use the change trajectory of unit processing time to represent the change of employee's fatigue level.

4.1 Shift-Related Data Analysis

The content of this section is how to estimate the standard operating time and judge the fatigue level under the consideration of some key factors, in order to provide a more accurate and practical reference value. Factors considered include shifts and gender.

As mentioned above, three shifts system is adopted with 8 h per shift. Three sets of data were analyzed and comparison, including rough face grinding (ProcCode97), fine face grinding (ProcCode99), and fine tapered face grinding (ProcCode118). The taken data is from May to July 2008 and we calculate the average of all employees' operating time per hour in each shift. Figure 4 shows the data analysis for the three shifts with the X-axis means time while Y-axis is operating time (min). Each dot in the chart represents the average of all actual processing times during this hour.

ProcCode99 and ProcCode118 have the same trend, while ProcCode97 is different. This might because the rough grinding process is simple and the difference

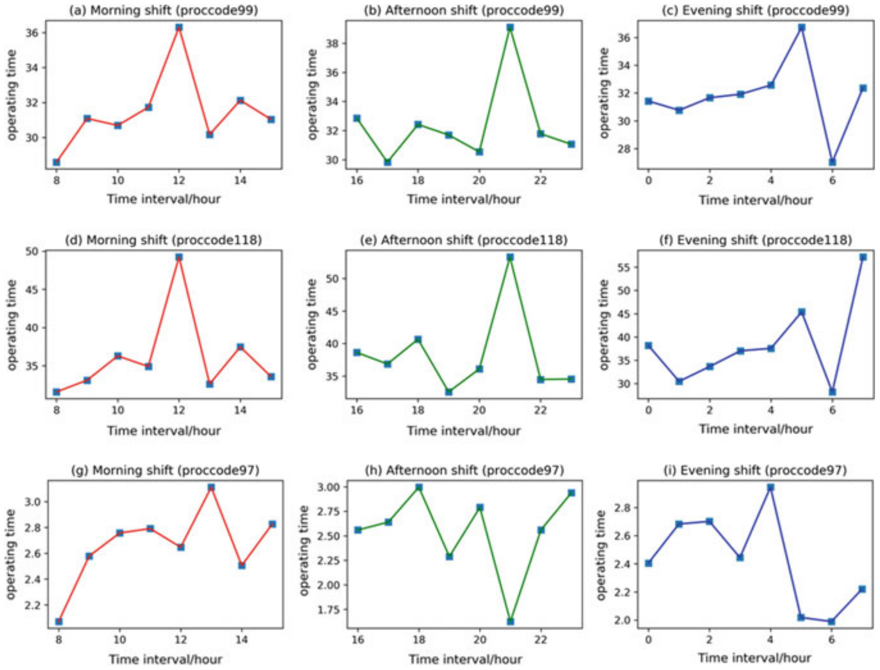


Fig. 4 Data set analysis with ProcCode of 99, 118 and 97

in unit processing time is basically within 1 min. There is no obvious difference and it cannot effectively represent the working status of employees.

Least squares cubic polynomial fitting is used to get the corresponding fatigue trend. The fitted graph of the second data set (ProcCode118) is shown in Fig. 5. Linear fitting, quadratic polynomial fitting, and cubic polynomial fitting are used to obtain the most suitable formula. It can be seen from Fig. 5 that the least square cubic polynomial fitting depicts the most realistic fatigue trajectory. The characteristic functions of the three curves are:

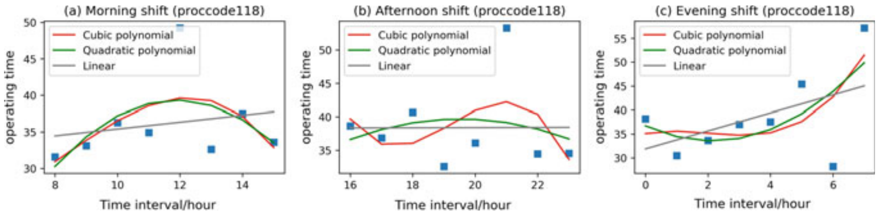


Fig. 5 Fitting curve of processing time

$$\begin{bmatrix} -0.06 & 1.58 & -10.18 & 43.84 \\ -0.29 & 16.86 & -321.20 & 2060.85 \\ 0.16 & -0.92 & 1.26 & 35.12 \end{bmatrix} \begin{bmatrix} x^3 \\ x^2 \\ x \\ 1 \end{bmatrix} = \begin{bmatrix} morningshift \\ afternoonshift \\ eveningshift \end{bmatrix}$$

The processing time for the morning shift peaked at 12 a.m. and dropped significantly after that. This shows that the morning shift workers have the highest accumulation of fatigue after four hours of work. During 12 a.m. to 1 p.m., there will be a certain lunch break, the fatigue of the workers has been significantly alleviated and the operating time then subsequently decreases. For afternoon shift, the fatigue level of employees reached its peak at 21:00, but at 22: 00–24: 00, which was generally considered to be very tired, the workers showed better vitality. Evening shift employees are in a relatively fatigued state for almost the entire working time and the fatigue level of night shift workers continued to rise with working hours but suddenly dropped at 6 a.m. This may be because in the summer, 6–7 a.m. is the most awake moment for a person. Sleepiness, the biggest challenge for night shift workers, has the weakest impact at this time.

Employee fatigue levels are simply divided into three levels and shown in Table 2. Corresponding management recommendations can also be implemented. Based on humane care and work hazard avoidance, the fatigue status of employees should be considered when performing workload distribution.

4.2 Gender-Related Data Analysis

To test the effect of gender on fatigue, we statistically analyzed 300 sets of data, including 26 employees, half of whom were women. Considering the possible impact of shifts on fatigue, we calculated the average processing time of male and female employees in 4-h intervals. Figure 6 depicts the corresponding analysis results. The average processing time for female and male employees was 33.72 min and 36.53 min, respectively. Although female employees perform better than male employees at any time interval, male and female employees show the same fatigue trend between 00:00 and 20:00. After 20:00, the fatigue of female employees is declining while that of male is just the opposite.

Let fatigue index α denote the fatigue level of male and female employees at different periods. The index α can be calculated through $\frac{\overline{t_i}-\overline{T_d}}{\overline{T_d}}, \overline{t_i} \in \overline{T}$. Index β indicates the change rate of employees' fatigue, $\beta = \frac{\overline{t_{i+1}}-\overline{t_i}}{\overline{t_i}}, \overline{t_i} \in \overline{T}$. The definitions of all variables are shown in Table 3.

Table 4 demonstrate the fatigue index of male and female employees throughout the day. The employee is in a fatigue state when α is positive and is energetic when α is negative. Index α is positively related to employees' fatigue level, the larger the index α, the higher the employee's fatigue level. Table 5 shows the change index of fatigue for employees during a day. Although both male and female workers have the

Table 2 Observation of the data

Shift	Time	Fatigue level	Management measures
Morning shift	08:00–10:00	No fatigue manifestation	Normal workload distribution based on standard operating times
	10:00–12:00	Moderate fatigue	Reduce workload slightly and appropriately
	12:00–13:00	Extreme fatigue	At this moment, the employee's sense of burnout and fatigue is at its peak, and the company can properly arrange the rest time and reduce the workload during this period
	13:00–14:00	No fatigue manifestation	The workload during this period can be increased appropriately
	14:00–16:00	Moderate fatigue	Normal workload distribution based on standard operating times
Afternoon shift	16:00–18:00	No fatigue manifestation	Normal workload distribution based on standard operating times
	18:00–19:00	Moderate fatigue	The fatigue level of employees has a small peak during this period. The company can arrange mealtimes and breaks and reduce the workload assessment during this period
	19:00–21:00	No fatigue manifestation	Normal workload distribution based on standard operating times
	21:00–22:00	Extreme fatigue	Schedule a break and reduce workload
	22:00–24:00	No fatigue manifestation	The workload during this period can be increased appropriately
Evening shift	00:00–05:00	Moderate fatigue	Reduce workload slightly and appropriately
	05:00–06:00	Extreme fatigue	Schedule a break and reduce workload
	06:00–07:00	No fatigue manifestation	Employees are relatively energetic at this moment, and the workload can be increased reasonably
	07:00–08:00	Moderate fatigue	Reduce workload slightly and appropriately

most rapid accumulation of fatigue in the interval I, the accumulation rate of male employees is slightly lower than that of female employees. In the second phase, there was a very significant recovery in the vitality of male workers, while female workers recovered only slightly. Since interval III, the fatigue accumulation rate of male employees showed a steady rise ($\beta = 0.03$). The fatigue symptoms of female employees subsided significantly in the interval V.

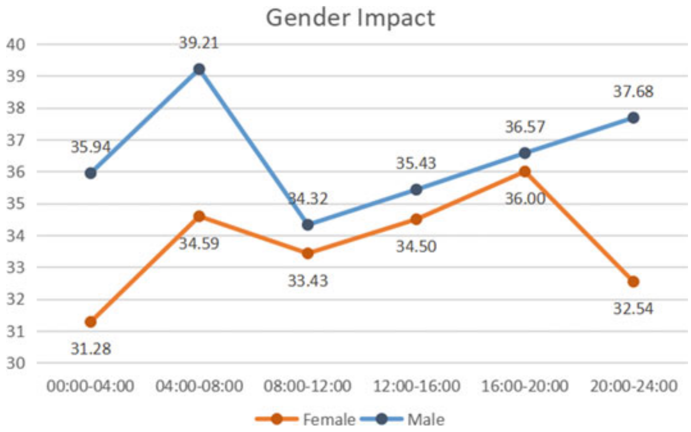


Fig. 6 The impact of gender on fatigue

Table 3 definitions of variables

Index	Definition
α	Fatigue index
β	Fatigue change index
$\overline{T}$	Average processing time of male and female employees in 4-hour intervals
$\overline{T_d}$	Average processing time of male and female employees during a day

Table 4 Employees' fatigue index throughout the day

	Period I	Period II	Period III	Period IV	Period V	Period VI
α (F)	−0.07	0.03	−0.01	0.02	0.07	−0.03
α (M)	−0.02	0.07	−0.06	−0.03	0.00	0.03

Table 5 Employees' fatigue change index during a day

	Interval I	Interval II	Interval III	Interval IV	Interval V
β(F)	0.11	−0.03	0.03	0.04	−0.10
β(M)	0.09	−0.12	0.03	0.03	0.03

By arranging work for male and female according to their different characteristics, work management can be performed more scientifically and work efficiency can also be improved. But how to plan effectively based on these characteristics requires further investigations.

5 Conclusions

This paper introduced an analytic approach for workers' fatigue examination using RFID-enabled production data. Numerous RFID-enabled data is classified based on different attributes after cleanse. Appropriate mathematical methods are used to obtain valid processing time data for analysis purposes. The actual unit processing time and fatigue are mapped to judge and analyze the influencing factors of fatigue. The collected data is transformed into effective management basis information. Both shifts and gender have corresponding effects on employees' fatigue levels. Employees have different fatigue accumulation processes in different shifts, and fatigue accumulation and vitality recovery are also different in men and women. Based on humane care and avoidance of workshop accidents, employee fatigue needs to be scientifically managed.

The contributions of this article can be draw as follows. Since RFID technology has been widely used in manufacturing companies' workshops to obtain real-time production data, the approach in this paper can be universally used. With the increasing popularity of humane care in society and enterprises, managers attach great importance to the physical and mental health of workers. Based on the impact of different factors on fatigue, employees' production and life can be allocated scientifically. Different workloads can be arranged in different shifts according to the corresponding fatigue state. Reasonable fatigue management measures such as planned short exercise or rest need to be arranged in the most sensitive period of fatigue to reduce the burden on employees. In addition, the different sense of fatigue brought about by gender should also be taken into account by managers. Workplace fatigue management should be controlled by both employees and managers, and both parties must recognize the negative effects of fatigue accumulation.

Further research could be carried out in three aspects. Firstly, the impact analysis of fatigue in the article is limited to one day. Longer-term effects of different factors on fatigue need further research. The impact within a month, a quarter, or a year can be analyzed in the future. Secondly, factors affecting workers' fatigue can be further expanded. In addition to shifts and gender factors, the impact of other factors (e.g., age) on fatigue is worth further exploration. In addition, the cross-effects of different factors on fatigue are worth considering. Finally, a work assignment system based on fatigue management for shop floor workers can be considered and established.

Acknowledgements We would like to acknowledge the support from Seed Fund for Basic Research in HKU (201906159001), The Knowledge Exchange impact project scheme in Faculty of Engineering, HKU (KE-IP-2019/20-31). Thanks for the data support from Huaiji Dengyun Auto-parts Holding Co Ltd.

References

1. K. Sadeghniiat-Haghighi, Z. Yazdi, Fatigue management in the workplace. Ind Psychiatry J. **24**(1), 12 (2015)
2. M. Ekstedt, M. Söderström, T. Åkerstedt, J. Nilsson, H.P. Søndergaard, P. Aleksander, Disturbed sleep and fatigue in occupational burnout. scandinavian J. Work, Environ. Health, 121–131 (2006)
3. A. Williamson, R. Friswell, Fatigue in the workplace: causes and countermeasures. Fatigue: Biomed. Health and Behav. **1**(1–2), 81–98 (2013)
4. M. Yung, Fatigue at the workplace: measurement and temporal development (2016)
5. M. Yung, P.L. Bigelow, D.M. Hastings, R.P. Wells, Detecting within-and between-day manifestations of neuromuscular fatigue at work: an exploratory study. Ergonomics **57**(10), 1562–1573 (2014)
6. S.E. Lerman, E. Eskin, D.J. Flower, E.C. George, B. Gerson, N. Hartenbaum, S.R. Hursh, M. Moore-Ede, Fatigue risk management in the workplace. J. Occup. Environ. Med. **54**(2), 231–258 (2012)
7. Z.S. Maman, M.A.A. Yazdi, L.A. Cavuoto, F.M. Megahed, A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. Appl Ergon **65**, 515–529 (2017)
8. C. Griffiths, J. Bowen, A. Hinze, Investigating wearable technology for fatigue identification in the workplace, in *IFIP Conference on Human-Computer Interaction,* (Springer, Cham 2017), pp. 370–380
9. S.M. Rajaratnam, C.B. Jones, Lessons about sleepiness and driving from the Selby rail disaster case: R v Gary Neil Hart. Chronobiol. Int. **21**(6), 1073–1077 (2004)
10. S. Folkard, D.A. Lombardi, Modeling the impact of the components of long work hours on injuries and “accidents”. Am. J. Ind. Med. **49**(11), 953–963 (2006)
11. P. Greig, R. Snow, Fatigue and risk: are train drivers safer than doctors? BMJ **359**, j5107 (2017)
12. W.J. Horrey, Y.I. Noy, S. Folkard, S.M. Popkin, H.D. Howarth, T.K. Courtney, Research needs and opportunities for reducing the adverse safety consequences of fatigue. Accid. Anal. Prev. **43**(2), 591–594 (2011)
13. K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication* (Wiley, 2010)
14. R.Y. Zhong, Q.Y. Dai, K. Zhou, X.B. Dai, Design and Implementation of DMES Based on RFID, in *2008 2nd International Conference on Anti-counterfeiting, Security and Identification,* pp. 475–477. IEEE
15. Q. Dai, Y. Liu, Z. Jiang, Z. Liu, K. Zhou, J. Wang, Mes wireless communication networking technology based on 433 mhz, in *2008 2nd International Conference on Anti-counterfeiting, Security and Identification*, pp. 110–113. (IEEE, 2008)
16. B. Yan, S. Shi, B. Ye, X. Zhou, P. Shi, Sustainable development of the fresh agricultural products supply chain through the application of RFID technology. Inf. Technol. Manage. **16**(1), 67–78 (2015)
17. T.M. Choi, W.K. Yeung, T.E. Cheng, X. Yue, Optimal scheduling, coordination, and the value of RFID technology in garment manufacturing supply chains. IEEE Trans. Eng. Manage. **65**(1), 72–84 (2017)
18. S.K. Kwok, K.K. Wu, RFID-based intra-supply chain in textile industry. Ind Manage. Data Syst. (2009)
19. C.K.H. Lee, K.L. Choy, G.T. Ho, K.M.Y. Law, A RFID-based resource allocation system for garment manufacturing. Expert Syst. Appl. **40**(2), 784–799 (2013)
20. R.Y. Zhong, G.Q. Huang, Q.Y. Dai, T. Zhang, Mining SOTs and dispatching rules from RFID-enabled real-time shopfloor production data. J. Intell. Manuf. **25**(4), 825–843 (2014)

Port Logistics Demand Forecast Based on Grey Neural Network with Improved Particle Swarm Optimization

Ruiping Yuan, Hui Wei, and Juntao Li

Abstract In order to improve the accuracy of port logistics demand prediction, the improved Particle Swarm Optimization algorithm, Grey Model and Neural Network are combined to construct an Improved Particle Swarm Optimization Grey Neural Network(IPSO-GNN) prediction model, in which the improved Particle Swarm Optimization algorithm is used to find the weight and threshold of the Grey Neural Network to improve the accuracy of the prediction. Using the logistics demand data of Dalian Port, the prediction effect of the proposed IPSO-GNN model is compared with that of the BP Neural Network model, the Grey model, the Grey Neural Network model and the standard Particle Swarm Optimization Grey Neural Network model. The empirical results show that the IPSO-GNN model has high precision and strong stability, which can predict port logistics demand effectively.

Keywords Improved particle swarm optimization grey neural network · Grey relational analysis · Port logistics demand prediction

1 Introduction

With the rise of northeast Asia economic circle, Dalian port logistics will meet greater development prospects and challenges. Port logistics demand is an important indicator in the port logistics system. Accurate prediction of port logistics demand will provide an important basis for the development of port logistics and logistics infrastructure planning. However, port logistics is a complex nonlinear system, which is

R. Yuan · H. Wei (✉) · J. Li
School of Information, Beijing Wuzi University, Beijing, China
e-mail: 18137873889@163.com

R. Yuan
e-mail: angelholyping@163.com

J. Li
e-mail: ljtletter@126.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_10

influenced by social, economic, natural and other factors, and the mapping relationship between these factors cannot be described in an accurate mathematical language, which leads to the prediction difficulty.

At present, the methods used in logistics demand forecasting can be divided into two categories: qualitative forecasting and quantitative forecasting. Qualitative prediction methods mainly include expert investigation method, Delphi method, subjective probability method and so on. Qualitative forecasting method is more flexible and simpler, but it is difficult to accurately describe the logistics demand due to the influence of subjective factors. Quantitative prediction methods mainly include moving average method, regression analysis method, exponential smoothing method, grey theory model, neural network prediction model and so on. Using the GM (1, 1) model to predict the port logistics demand of Guangxi Beibu gulf, providing decision-making basis for relevant government departments [1]. Establishing the BP neural network model to predict the port logistics demand of Cao Feidian port and put forward policy Suggestions for the future development of modern logistics [2]. Considering the characteristic of logistics demand with nonlinear changes, proposing a combined forecasting model based on BP and RBF neural network, the empirical results show that the combination forecast model than single prediction model has higher prediction accuracy, reducing the probability of error effectively in larger, making the forecast results closer to reality, and propose the port logistics development planning for the future [3]. Neural network can be used to solve nonlinear and complicated prediction problems, but it has some disadvantages such as large amount of data required for training, high cost of data collection, and easy occurrence of local optima. Although the grey model requires only a small amount of data, the prediction accuracy is not high. The Grey Neural Network (GNN) model makes up for the above defects to some extent, combining the advantages of strong nonlinear fitting ability of Neural Network and high accuracy of Grey Model under the condition of small amount of calculation and few samples. The Grey BP Neural Network model has higher prediction accuracy than the single grey prediction model [4, 5]. The grey neural network model could help enterprises predict the market demand better after transportation interruption, and then empirical research tested its possibility [6]. However, due to the randomness of the determination of initial weight and threshold of GNN, the network is prone to fall into the local optimal, the results of each prediction are different and the deviation is large. Particle Swarm Optimization (PSO) is an optimization algorithm based on swarm intelligence theory, which has good robustness and global search ability. Optimization of GNN weights and thresholds by PSO can make up for the above shortcomings. Particle swarm optimization grey neural network model to predict proton exchange membrane fuel cell (PEMFC) degradation, and the results showed that the prediction accuracy of this method was high [7].

Because the port logistics demand data is relatively small and has strong nonlinear relationship, it is very important to establish a suitable model to predict the future port logistics demand, according to the port logistics demand data. GNN has the advantages of strong non-linear fitting ability, small calculation amount and high calculation accuracy for small sample data. PSO has good robustness and global search ability, which is fully suitable for port logistics demand prediction. However, the standard PSO is easy to fall into local minimum point and the problem such as premature convergence, so this article on the basis of the above study, the improved PSO algorithm (IPSO) optimizes geri weis-corbley weights and thresholds by the weights of the nonlinear regressive strategy and strategy of dynamic accelerated learning factor, adjusting and balancing between global and local search capabilities, building the improved particle swarm optimization of grey neural network forecast model, and applied to port logistics demand forecasting in order to improve the prediction accuracy.

2 Prediction Model Based on IPSO-GNN

2.1 Gray Neural Network

The principle of GNN is to map the whitening equation of Grey Model (GM) into a BP neural network. When the network is trained and converges, the corresponding connection weight coefficient $a, b_1, b_2, \ldots, b_{n-1}$ is extracted from the trained network to obtain a whitening differential equation, and then the data fitting and prediction are performed according to this differential equation [8, 9]. The topology of the gray neural network is shown in Fig. 1.

In the figure, t is the serial number of the input parameter, $y_2(t), y_3(t), \ldots, y_n(t)$ are the network input parameter, $\omega_{21}, \ldots, \omega_{3n}$ are the network weight, $\omega_{11} = a, \omega_{21} = -y_1(0)$,$\omega_{22} = \frac{2b_1}{a}, \ldots$,$\omega_{2n} = \frac{2b_{n-1}}{a}$,$\omega_{31} = \omega_{32} = \cdots = \omega_{3n} = 1 + e^{-at}$, y_1 is the network prediction value, and $\theta = \left(1 + e^{-at}\right)(d - y_1(0))$ is the output node threshold of LD layer. The error criterion of the algorithm adjusts the weight

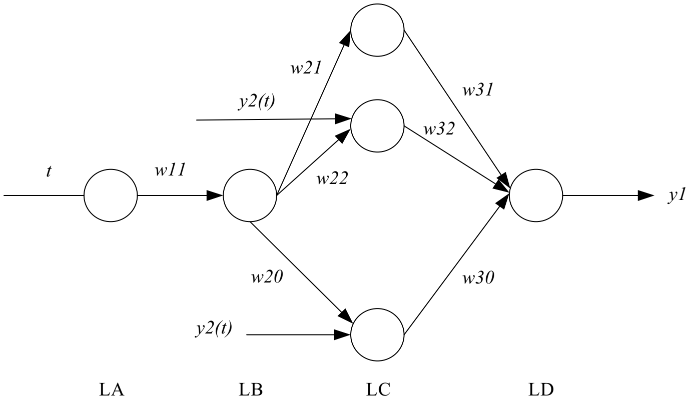


Fig. 1 The topology of the grey neural network

to minimize the deviations. Therefore, it is crucial to obtain the initial value of $a, b_1, b_2, \ldots, b_{n-1}$ of GNN.

2.2 Improved Particle Swarm Optimization Algorithm

The basic idea of PSO [10, 11] is to use the information contained in each particle to express the optimal solution of the optimization problem. The important parameters affect the performance of PSO including flight speed, inertia weight and learning factor. Therefore, optimizing the inertia weight ω and learning factor can improve the performance of the algorithm. Inertia weight ω is used to control the exploration and convergence ability of the algorithm. When the ω is larger, it has stronger global search ability, it is more conducive to local search when ω is smaller. Learning factor is used to guide the algorithm to search for the optimal solution. In the early stage of optimization, particles should be encouraged to search in a larger space to keep the diversity of particles. Therefore, c_1 is larger and c_2 is smaller to speed up the search speed of particle swarm. In the later stage of optimization, the center of gravity of particle swarm is in the global optimal solution, so c_1 is smaller and c_2 is larger to keep particle swarm accurate search. Based on the algorithms proposed in literature [12–14], this paper firstly adopts an improved weight nonlinear descending strategy for setting the inertial weight ω, so that the PSO algorithm can better balance the global and local search capabilities. Secondly, adopting the learning factor strategy of dynamic acceleration, c_1 and c_2 change linearly with the number of iterations. The improved particle swarm optimization algorithm is shown as follows:

- In this paper, inertial weight ω can better balance and adjust the local and global search capability of PSO algorithm by introducing tangent function. The calculation formula is as follows:

$$\omega = \omega_{\min} + (\omega_{\max} - \omega_{\min}) * \tan\left(\frac{\pi}{4} * \left(1 - \frac{t}{T_{\max}}\right)^{\theta}\right) \tag{1}$$

- The learning factor c_1 and c_2 of dynamic acceleration in this paper are between the minimum value and the maximum value all the time, so it is easy to avoid c_1 and c_2 falling into the local optimum if the learning factor is too large or too small. The calculation formula is as follows:

$$c_1(t) = c_1\mathrm{max} - \left(c_1\mathrm{max} - c_{1\mathrm{min}}\right) * \frac{t}{T_{\max}} \tag{2}$$

$$\mathrm{c}_2(\mathrm{t}) = \mathrm{c}_{2\mathrm{min}} + \left(\mathrm{c}_2\mathrm{max} - \mathrm{c}_{2\mathrm{min}}\right) * \frac{\mathrm{t}}{\mathrm{T}_{\mathrm{max}}} \tag{3}$$

Among them, t is the current iteration number, $T_{\max}$ is the maximum iteration number, θ is the curve adjustment factor, and θ is 2 in this algorithm after several experiments. In general, $\omega_{\max} = 1$, $\omega_{\min} = 0.5$, the values of $c_{1\min}$ and $c_{2\min}$ are 1, and the values of $c_{1\max}$ and $c_{2\max}$ are 2.

The updated formula of particle velocity and position of the improved particle swarm optimization algorithm are as follows:

$$v_{id}(t+1) = v_{id}(t) + c_1(t)r_1(p_{id} - x_{id}(t)) + c_2(t)r_2\big(p_{gd} - x_{id}(t)\big) \tag{4}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{5}$$

Among them, r_1 and r_2 is the random number [0, 1], x_{id} is the current position of the particle, p_{id} is the optimal solution of the particle history, and p_{gd} is the global optimal solution.

2.3 Improved Particle Swarm Optimization Algorithm to Optimize Grey Neural Network

GNN has a fast convergence speed, but its initial weights and thresholds are determined randomly, so that the network is prone to local optimization [14]. Therefore, Improved Particle Swarm Optimization algorithm can obtain the optimal GNN parameters. IPSO-GNN algorithm uses IPSO algorithm to optimize the $a, b_1, b_2, \ldots, b_{n-1}$ parameters of GNN, and takes the mean square error of training samples as the function of individual adaptive value. The global optimal solution as the initial weight and threshold of GNN to train the network, so that achieve the training goal of the network. The specific algorithm steps are as follows:

(1) Divide the original data into training samples and test samples;

(2) *Initialization parameters:* initialize the maximum and minimum value of the learning factor of the particle, the maximum and minimum value of the inertial weight, the position and speed of the particle, and assign a set of parameters corresponding to each position the initial particle passes through. Since the initial weight and threshold of GNN are determined by equal n parameters, the dimension of the particle swarm is $D = n$.

(3) Set the fitness function of particle swarm, prediction mean square error (MSE) of the training sample as the function of individual adaptive value. Because there is only one output neuron of GNN, formula can be simplified as: $MSE = \frac{1}{n}\sum_{i=1}^{n}(o_i - y_i)^2$, among them, n as the number of training samples, o_i is the actual output of the i th sample, y_i is the expected output of i th sample;

(4) Compare and analyse the fitness value of each particle and its corresponding optimal value, and then judge whether satisfying the condition of iteration. If so, these are the optimal parameter combination. Go to step (6); otherwise, go to step (5);

(5) According to formula (4) and (5), the particle speed and position are repeatedly update, and judge if meet the optimal solution conditions. When meeting the conditions that the minimum precision value of fitness or the maximum number of iterations, go to step (6); otherwise, go to step (5);

(6) Obtain the optimal parameters;

(7) Substitute the optimal parameters as the initial weight and threshold of GNN into the network for training until the training error (or iteration times) of the network reaches the predetermined value. The IPSO-GNN algorithm flow shows in Fig. 2.

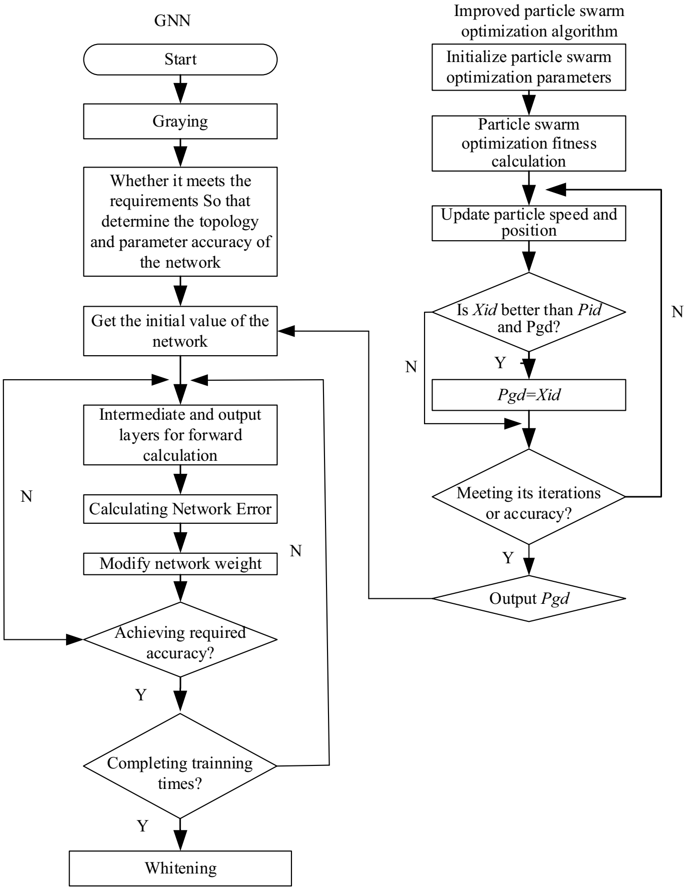

GNN
Start
Graying
Whether it meets the requirements So that determine the topology and parameter accuracy of the network
Get the initial value of the network
Intermediate and output layers for forward calculation
Calculating Network Error
Modify network weight
Achieving required accuracy?
N
Y
Completing trainning times?
N
Y
Whitening
Improved particle swarm optimization algorithm
Initialize particle swarm optimization parameters
Particle swarm optimization fitness calculation
Update particle speed and position
Is Xid better than Pid and Pgd?
N
Y
Pgd=Xid
N
Meeting its iterations or accuracy?
Y
Output Pgd

Fig. 2 The IPSO-GNN algorithm flow

3 Selection of Logistics Demand Forecast Index of Dalian Port

3.1 Influencing Factors Analysis of Dalian Port Logistics Demand Forecast

Zhu et al. Select four indicators that are the added value of the primary industry, the added value of the secondary industry, the total amount of import and export, and the investment in fixed assets of the whole society as the indicators to predict the port cargo logistics demand [1]. Gao et al. select the port city GDP, industrial GDP, tertiary industry GDP, total foreign trade, total retail sales, per capita income, and per capita consumption level as the indicators affecting the port logistics demand of Feidian [2]. Cai and Huang Select the three industrial values, total imports and exports, total retail sales of consumer goods and fixed asset investment in the direct economic hinterland as the influencing factors of the logistics demand of Shantou port [3]. Basing on relevant literature, because the port logistics demand amount is closely related to the hinterland economic aggregate, which can predict port logistics demand rely on the correlation between the two. combined with the actual situation of Dalian port logistics development, this paper selects the total regional GDP (a_1), added value of primary industry (a_2), added value of secondary industry (a_3), added value of tertiary industry (a_4), fixed asset investment of the whole society(a_5), total retail sales of consumer goods (a_6), total imports and exports (a_7), annual disposable income (a_8), and annual per capita consumption expenditure (a_9) nine indicators to predict the cargo logistics demand in Dalian Port. The statistical data of logistics demand impact indicators of Dalian port in 2001–2018 is shown in Table 1. Among them, the unit of a_1, a_3, a_4 and a_6 are RMB 100 million, the unit of a_5 and a_7 are dollars 100 million, the unit of a_8 and a_9 are RMB, the unit of a_0 is 100 million tons.

3.2 Analysis and Selection of Logistics Demand Index of Dalian Port

Grey relation analysis is based on the grey system theory proposed by Professor Julong Deng. Processing the various factors through data in incomplete information to find the correlation degree among them [15–17]. Introducing grey correlation analysis method is to further determine the correlation among the logistics demand of Dalian port and various impact indicators. The statistical data of the logistics demand impact indicators of Dalian port is shown in Table 1. It is necessary to select the indicators that have a significant impact on the logistics demand of Dalian port, namely the key indicators (the correlation degree is greater than 0.6). The correlation degree among each indicator and the logistics demand of Dalian port is determined by Calculated by DPS software, the specific results are as follows:

Table 1 Statistics of logistics demand impact indicators of Dalian port in 2001–2018

Year	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_0
2001	1235.6	111.4	574.2	550	305.1	534.2	112.4	7418	6512	1.05
2002	1406	118.4	661.1	626.5	367.9	591.9	129.86	8200	7118	1.12
2003	1632.6	145.6	782.1	704.9	506.9	568.45	155.4	9101	7760	1.26
2004	1961.8	153.1	983.3	825.4	716.2	645.2	194.3	10,378	8672	1.45
2005	2150	185.2	994	970.8	1110.5	732	235.23	11,994	9996	1.71
2006	2569.7	208.6	1229	1132	1469.5	839.3	293.24	13,350	10,534	2
2007	3131	247.6	1536.5	1346.9	1930.8	983.3	363.03	15,109	12,135	2.2
2008	3858.2	289.1	1993.9	1575.2	2513.4	1182.6	449.09	17,500	14,101	2.46
2009	4417.7	313.4	2314.8	1789.5	3273.5	1396.7	403.47	19,014	15,330	2.73
2010	5158.1	345.1	2645.5	2167.5	5084.3	1639.8	501.95	21,293	16,580	3.14
2011	6150.1	395.7	3204.2	2550.2	4580.1	1924.8	585.25	24,276	18,846	3.4
2012	7002.8	451.4	3634.8	2916.7	5654.1	2224	625.63	27,539	20,417	3.7
2013	7650.8	477.6	3892	3281.2	6478.1	2526.5	676.55	30,238	22,516	4.1
2014	7655.6	441.8	3696.5	3517.2	6773.6	2828.4	645.78	33,591	27,482	4.2
2015	7731.6	453.3	3580.8	3697.5	4559.3	3084.3	550.91	35,889	25,824	4.15
2016	6810.2	462.8	2849.9	3497.6	1436.4	3410.1	547.6	38,050	27,119	4.4
2017	7363.9	477.1	3052.6	3834.3	1652.8	3722.5	666.43	40,587	27,191	4.6
2018	7668.5	442.7	3241.6	3984.2	1819.7	3880.1	759.06	43,550	29,928	4.7

Note The data are from the website of Dalian Statistics Bureau

$$\rho_{a1} = 0.6597\rho_{a2} = 0.7207\rho_{a3} = 0.5616$$
$$\rho_{a4} = 0.6578\rho_{a5} = 0.3838$$

$$\rho_{a6} = 0.5823\rho_{a7} = 0.6475\rho_{a8} = 0.6406$$
$$\rho_{a9} = 0.6793$$

The larger the correlation value, the greater the impact of this index on the logistics demand of Dalian port is. By sorting the above correlation values, we can get: the values of a_2, a_4, a_9, a_1, a_7 and a_8 six variables are relatively large, that is to say, the added value of the primary industry, the added value of the tertiary industry, the annual per capita consumption expenditure, the GDP of the whole region, the total amount of import and export, and the annual disposable income six indicators are selected as the key indicators. The logistics demand forecast index set of Dalian port is shown in Table 2.

Table 2 Dalian port logistics demand forecast index set

Dalian port logistics demand forecast index	Economic scale index	GDP of the whole region
	Index of industrial organization	Added value of primary industry and tertiary industry
	Domestic and foreign trade indicators	Total imports and exports
	Consumption index of residents	Annual disposable income and annual per capita consumption expenditure

4 Logistics Demand Prediction Dalian Port Based on IPSO-GNN

4.1 Establish a Prediction Model Based on IPSO-GNN

Take the data in Table 1 from 2001 to 2015 as the network training data, and the data in 2016–2018 as the network test data. In order to prevent the net input absolute value too large cause the saturation of neuron output, so that reduce the convergence of training network, it is necessary to make Dalian port logistics demand indicators statistics normalization. IPSO-GNN model sets six input variables and one output variable, and network hidden layer adopts multi-layer deep learning mode.

4.2 Model Prediction Results and Analysis

Empirical research adopts the gradual model analysis method, including BP neural network model, GM (1, 1) model, GNN model, standard particle swarm optimization grey neural network (PSO-GNN) model and IPSO-GNN model to predict the logistics demand of Dalian port. Using matlabr2012b programming software realize the above algorithm. The calculation results of prediction error and mean square error of each prediction model are shown in Table 3, in which the unit of actual values and prediction values are 100 million yuan.

From the Table 3, the following conclusions can be drawn:

- From the comparison of real and predicted values in 2016–2018, the mean error of prediction value under BP neural network model, grey GM (1,1) model, GNN model, PSO-GNN model and IPSO-GNN model are 11.12%, 24.9%, 5%, 2.96% and 1.71% respectively, the mean square errors are 0.2622, 1.4138, 0.0582, 0.0184 and 0.0076. The IPSO-GNN prediction model proposed in this paper has the smallest deviation and the highest accuracy from the original data, which is suitable for the prediction of port logistics demand.

Table 3 Comparison of simulation results

Year	Actual value	BP		GM (1, 1)		GNN		PSO-GNN		IPSO-GNN	
		Prediction value	Relative error (%)	Prediction value	Relative error (%)	Prediction value	Relative error (%)	Prediction value	Relative error (%)	Prediction value	Relative error (%)
2016	4.4	3.969	9.8	5.19	17.7	4.188	4.83	4.239	3.66	4.534	3.04
2017	4.6	4.05	11.96	5.694	23.79	4.279	6.99	4.485	2.51	4.659	1.29
2018	4.7	4.154	11.61	6.261	33.22	4.551	3.17	4.573	2.7	4.737	0.79
Mean error (%)		11.12		24.9		5		2.96		1.71	
MSE		0.2622		1.4138		0.0582		0.0184		0.0076	

- Although there are random fluctuations in the original port logistics demand, the mean error and mean square error of the model in this paper are the smallest. With the increase of the forecast year, the relative error of the forecast results than other forecast models have volatility, while the relative error of the forecast results of the algorithm in this paper decreases year by year, with high stability. The IPSO-GNN model prediction results are in good agreement with the original port logistics demand data, which proves the stability of the algorithm.

5 Conclusions

Based on the study of particle swarm optimization and grey neural network, combined with the characteristics of port logistics demand forecasting, this paper applies innovatively the improved particle swarm optimization grey neural network model to port logistics demand forecasting. The improved particle swarm algorithm optimizes the weight and threshold value of the grey neural network to improve the prediction accuracy, which effectively solves the problems that randomness given by the initial parameters of the grey neural network and easy to fall into local optimum. This present paper compares with BP neural network model, GM (1, 1) model, GNN model, PSO-GNN model, IPSO-GNN model in Dalian port logistics demand prediction results, shows that IPSO-GNN prediction model can significantly improve the prediction accuracy of port logistics demand, and the model stability is stronger.

References

1. N. Zhu, D. Chen, C. He, Based on the gray GM (1, N) model of Guangxi Beibu gulf port logistics prediction research. J. Math. Pract. Underst. **47**(23), 303–310 (2017)
2. X. Gao, J. Xu, Y. Gao, Port logistics demand prediction research based on BP model. J. Logist. Technol. **33**(3), 99–101 (2014)
3. W. Cai, H. Huang, Based on the combination of BP and RBF neural network model to predict the port logistics demand study. J. Zhengzhou Univer. (Eng. Sci.) **40**(5), 85–91 (2019)
4. A. Wang, Y. Liu, Human resource demand forecasting method based on grey BP neural network model. Statist. Decis.-Making, **34**(16), 181–184 (2008)
5. X. Liu, B. Moreno, A. Salomé García, A grey neural network and input-output combined forecasting mode primary energy consumption forecasts in Spanish economic sectors. Energy **115**(11), 1042–1054 (2016)
6. T.S. Liu, S. Chen, An improved grey neural network model for predicting transportation disruptions. Expert Syst. Appl. **45**(3), 331–340 (2016)
7. K. Chen, S. Laghrouche, A. Djerdir, Degradation prediction of proton exchange membrane fuel cell based on grey neural network model and particle swarm optimization. Energ. Convers. Manage. **195**(9), 810–818 (2019)
8. Liu, T.S., S. Chen, An improved grey neural network model for predicting transportation disruptions. Expert Syst. Appl. **45**(C), 331–340 (2015)
9. H.T. Lei, X. Xu, Based on improved particle swarm optimization algorithm of railway freight volume forecasting of grey neural network. J. Comput. Appl. **32**(10), 2948–2951 + 2962 (2012)

10. X. Tang, G. Qiu, Y. Li, Based on particle swarm optimization and SOM network clustering algorithm research. J. Huazhong Univ. Sci. Technol. (Nat Sci Ed) **35**(5), 31–33 + 37 (2007)
11. X. Wei, H. Pan, *Particle Swarm Optimization and Intelligent Fault Diagnosis* (National defence industry press, Beijing, 2010)
12. F. Zhou, Y. Lv, L. Shi, Improved particle swarm algorithm to optimize the grey neural network forecast model and its application. J. Stat. Decis. **32**(11), 66–70 (2017)
13. R. Xu, Y. Wang, F. Wang, Based on improved PSO and BP algorithm express traffic prediction. J. Comput. Integr. Manuf. Syst. **24**(7), 1871–1879 (2018)
14. Z. Zhao, F. Yang, Z. Zhang, The particle swarm algorithm to optimize the grey neural network satellite clock error prediction. J. Navig. Positioning **6**(2), 53–56 + 81 (2018)
15. S. Liu, H. Cai, Y. Yang, Research progress of grey relational analysis model. Syst. Eng. Theory Pract. **33**(8), 2041–2046 (2013)
16. X. Li, Agricultural products logistics demand forecast based on grey linear combination model. J. Beijing Jiaotong Univ. (Soc. Sci. Ed.) **16**(1), 120–126 (2017)
17. J. Liao, C. Lin, Optimization and simulation of job-shop supply chain scheduling in manufacturing enterprises based on particle swarm optimization. Int. J. Simul. Model. **18**(1), 187–196 (2019)

A Review of the Schedulability Tests and Optimal Design for Rate-Monotonic Scheduling

Yang Li, Xiuli Wang, Tianying Liu, Youwei Wang, Jianming Zhu, and Meijiao Duan

Abstract The schedulability analysis of Rate-Monotonic scheduling algorithm is still under active investigation by researchers. New dimensions have been explored accordingly ranging from faster schedulability tests to optimal design of real-time systems. Recently, researchers extended the schedulability problem to even a more generalized problem: Rate-Monotonic Optimal Design Problem (RM-ODP) in which the execution time is limited in interval rather than a sole point. The RM-ODP deals with adjusting the execution times of tasks such that (i) the system is RM schedulable, and (ii) certain system performance (e.g. processor utilization) is optimized. In this paper, we first evaluate existing relevant feasibility tests by categorizing them into two broader classes: inexact and exact conditions. As a second contribution, we comprehensively review the RM-ODP and summarize associated advantages and disadvantages. Some recommendations on the schedulability tests are provided in order to assist system designers in selecting the most appropriate technique for the optimal design of the system to be run under RM scheduling policies.

Keywords Real-time systems · Rate-monotonic scheduling · Schedulability tests · Optimal design

Y. Li · X. Wang · T. Liu · Y. Wang · J. Zhu · M. Duan (✉)
School of Information, Central University of Finance and Economics, Beijing, China
e-mail: duanmeijiao@cufe.edu.cn

Y. Li
e-mail: liyang@cufe.edu.cn

X. Wang
e-mail: wangcufe@163.com

T. Liu
e-mail: 2019212039@email.cufe.edu.cn

Y. Wang
e-mail: ywwang15@126.com

J. Zhu
e-mail: zjm@cufe.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_11

1 Introduction

In real-time systems, the correctness of computation not only depends on the logic result but also on the timing constraints. Real-time scheduling is one of the research topics in real-time systems and has been widely investigated in the past forty years. Real-time scheduling algorithms ensure the constraints of timing, temporal and sources are satisfied by determining which processor a task is to be executed on and assigning relevant resources under limited system resources (e.g. CPU).

In 1973, Liu and Layland proposed the Rate-Monotonic (RM) algorithm for the scheduling of preemptive tasks with fixed priorities in uniprocessor [16]. After Liu and Layland's seminal work, many researchers have developed test algorithms to determine whether a task set is feasibly schedulable in RM. In sufficient conditions (also referred to as inexact tests), The LL-bound is raised by the literatures including [3, 18] calculating the utilization upper bound by using task periods and execution times directly. [4, 12, 13] presented the tests based on harmonic chain method efficient for the task periods with multiple relationships. Construction methods were proposed in papers [4, 10] where the original task set is transformed into a new task set of which the schedulability can be easily determined, and the original task set is schedulable if the transformed one is schedulable. Authors of [14] developed the utilization upper bound by solving linear programming problems. In 1989, Lehoczky et al. [15] developed the first exact tests by checking the satisfibility of a list of inequalities on the scheduling point set with finite elements. The scheduling point set was later reduced by [21]. These methods are categorized as the Scheduling Points Tests. Audsley et al. [1] introduced the exact schedulability test based on Response Time Analysis (RTA) which is an iterative method. Papers [19] improved the RTA by constructing the initial values. Moreover, the RTA were also applied to the inexact tests presented in [11, 22].

In this paper, we briefly review two areas of the Rate-Monotonic scheduling theory. The first part is the decision algorithms for the schedulability in RM. This part consists of both the inexact and exact tests existing in literatures. The second part is the Rate-Monotonic Optimal Design Problem (or C-AP) extended from the schedulability tests just answering the simple answer "YES or NO".

2 Inexact Schedulability Tests

2.1 Utilization Upper Bound Method

1. *Liu and Layland (LL) Bound Condition*

Theorem 1 (LL-Bound Condition) *A task set τ is schedulable in the RM algorithm if*

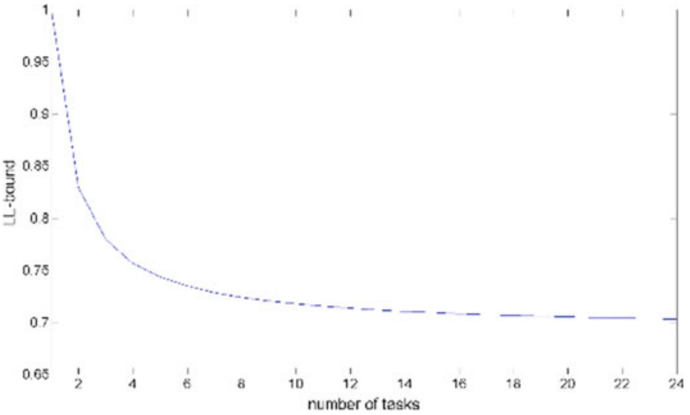


Fig. 1 Relationship between the task number n and the LL-bound $L(n)$

$$U \leq n\left(2^{1/n} - 1\right) = L(n) \tag{1}$$

where n denotes the number of tasks in τ.

The LL-bound function $L(n)$ monotonically decreases from $2\left(\sqrt{2} - 1\right) \approx 0.83$ when $n = 2$ to $\ln(2) \approx 0.69$ when n trends to the positive infinity. The relationship between the task number n and the LL-bound $L(n)$ is shown in Fig. 1.

2. *Hyperbolic Bound Condition*

Bini et al. introduced an efficient condition, the Hyperbolic Bound (HB) Condition for verifying the feasibility of scheduling tasks in Rate-Monotonic algorithm [3]. The HB condition is similar to the utilization oriented condition and is described in Theorem 2 [3].

Theorem 2 *Let* $\tau = \{\tau_1, \tau_2, \ldots, \tau_n\}$ *be a set of periodical tasks, where each task* τ_i *is characterized by a processor utilization* U_i. *Then* τ *is schedulable in the RM algorithm if*

$$\prod_{i=1}^{n} (U_i + 1) \leq 2 \tag{2}$$

The Hyperbolic Bound improves the LL-bound for any task set; the LL-bound condition implies the Hyperbolic Bound condition. Consider the task set satisfies the LL-bound condition, then

$$U = U_1 + U_2 + \cdots + U_n \leq n\left(2^{1/n} - 1\right)$$

and add one to each U_i, then the right hand side of the above inequality should be added by n:

$$(U_1+1)+(U_2+1)+\cdots+(U_n+1)\le n\bigl(2^{1/n}-1\bigr)+n=n2^{1/n}.$$

According to the AM-GM inequality, we have

$$n2^{1/n}\ge(U_1+1)+(U_2+1)+\cdots+(U_n+1)\ge n[(U_1+1)(U_2+1)\cdots(U_n+1)]^{1/n}.$$

Then the HB condition in (2) is obtained after simplifying the above inequality. When $U_1=U_2=\cdots=U_n$, the second inequality sign becomes an equal sign, and in which case, the HB condition is equivalent to the LL-bound condition.

2.2 Harmonic Chain Method

The Harmonic Chain method is firstly proposed by Kuo and Mok in their paper [12], which extends the LL-bound [16] by detecting harmonic relationships among task periods. A Harmonic Chain is defined as a list of numbers (task periods) wherein each number divides every number after it [4]. This method can effectively deal with the periods with multiple relationships. Kuo and Mok developed the harmonic chain (HC) schedulability condition [12], which improves the LL-bound by replacing the task number n with the size of the harmonic base of the task set. Chen et al. [4] introduced a strategy to compute a higher bound based on the harmonic chain condition with higher efficiency. Kuo et al. in paper [13] proposed the root condition based on the harmonic chain condition, which also improves the accept ratio of the HC condition.

A Harmonic Chain is obtained by finding the harmonic base of a task set. The harmonic base and harmonic chain are defined in the following [12].

Definition 1 (Harmonic Base and Harnomic Chain) Let S be the set of periods (positive numbers) of a task set τ. A subset R of S is said to be a harmonic base of the task set τ if there is a partition Γ of S into $|R|$ subset such that:

(1) each number of R is the smallest element in exactly one member of Γ, that is, the smallest number in each partition subset is in the set R, and
(2) if x and y are two elements in the same member of the partition Γ, then either x divides y or y divides x.

Moreover, each subset in the partition Γ is called a harmonic chain.

The harmonic chain condition can efficiently deal with the task periods with division relationships. Specifically, if all the periods (except the smallest one) are multiples of their previous periods in real-time system, which is called the *simply periodic system*, the schedulability bound can be raised to 1 according to the harmonic chain condition. However, the harmonic chain condition degrades to the LL-bound condition if all the task periods are relatively primes.

2.3 Construction Method

Han and Tyan introduced the following theorem for the schedulability test [10].

Theorem 3 Given a task set τ, if there exist another task set $\tau\prime$ such that $T_i\prime \leq T_i$ and $C_i\prime = C_i$, for $1 \leq i \leq n$, and $\tau\prime$ is schedulable by RM, then τ is also schedulable by RM.

Han and Tyan proposed two polynomial-time algorithms, Sr and DCT for finding such a task set $\tau\prime$. The two algorithms are originally devised for scheduling the real-time *distance-constrained task* set [18].

DCT Algorithm The DCT algorithm can be described as follows. For each f, $1 \leq f \leq n$, let $T_f\prime = T_f$, and then each task period T_i is transformed recursively. For each $i > f$, T_i is transformed to the largest integral multiple of $T_{i-1}\prime$ that is less than or equal to T_i, i.e.

$$T_i\prime = T_{i-1}\prime \left\lfloor \frac{T_i}{T_{i-1}\prime} \right\rfloor, \quad i = f+1, f+2, \ldots, n$$

Algorithm 2 DCT Algorithm

```
Input: task set $\tau$
Output: new task set $\tau' = \{(C_i, T_i') | 1 \leq i \leq n\}$
function DCT($\tau$)
    $\min_f = 1$
      $\min_U = +\infty$
         for $f = 1$ to $n$ do
            $Z_f = T_f$
            for $i = f + 1$ to $n$ do
                $Z_i = Z_{i-1} \lfloor T_i / Z_{i-1} \rfloor$
            end for
            for $i = f - 1$ downto $1$ do
                $Z_i = Z_{i+1} \lfloor Z_{i+1} / T_i \rfloor$
            end for
            $U = \sum_{i=1}^{n} C_i / Z_i$
             if $U < \min_U$ then $\min_U = U$  $\min_f =$
                for $i = 1$ to $n$ do
                    $T_i' = Z_i$
                end for
            end if
          end for
          return $\tau'$
    end function
```

2.4 Response Time Analysis

RTA is an iterative method to determine the schedulability of a task set in RM, and it is initiated by Harter in 1984, who used a temporal logic proof system to derive the Time Dilation Algorithm [21].

Audsley et al. presented the method for calculating the response time with fix-point iteration and developed an exact schedulability test [1]. Since then RTA has been extended to deal with more complex cases in the schedulability, such as sporadic processes, kernel overheads, blocking by low priority processes, release jitter, arbitrary large deadlines, etc.

Response time analysis can be used to develop both the exact and inexact (sufficient but not necessary) tests for the schedulability in RM. The exact tests are introduced in Sect. 3.

Han et al. proposed a sufficient schedulability condition based on the response time analysis [11]. We refer to this condition as Han-RTA. For each task τ_i the worst case response time is selected from a set of ratios $[T_i/T_k]$ $(1 \le k \le i)$. Han-RTA condition is defined in the following theorem [11].

Theorem 4 *Given a task set τ, let $V_{i.k} = [T_i/T_k]T_k$ for tasks τ_i and τ_k. The task set τ, for all $\tau_i \in \tau$, is schedulable if there exists k $(1 \le k \le i)$ satisfying that*

$$C_i + \sum_{j=1}^{i-1} \left[V_{i,k}/T_j \right] C_j \le V_{i,k} \tag{3}$$

The time complexity of the RTA-based test is $O\left(n^3\right)$ in the worst case. The simulation experiment in [11] shows that Han-RTA is very close to the exact schedulability test.

2.5 Summary of the Inexact Tests

Our analysis show that DCT condition is recommended for online systems, and the literature [7] also shows that DCT condition performs best in general online mode. However, DCT is unable to handle situations subject to constrained circumstances i.e. runtime constraints etc. Hence, we recommend choosing a schedulability condition should be based on two things: (i) the time complexity of the test, and (ii) the ability of the test to handle the task set parameters such as task period and execution time.

Another aspect of inexact test that is worth exploring is the acceptance ratio of the test, where higher acceptance ratio is desirable. Unfortunately, as highlighted in literature [7], acceptance ratio is closely associated with time complexity of the schedulability conditions i.e. higher acceptance ratio requires higher time complexity. Based on the nature of the system, a designer may chose DCT for higher acceptance ration with $O\left(n^2\right)$. In cases, when $O\left(n^2\right)$ is not affordable we suggest the following three test in decreasing order of their time complexities and acceptance ratios: T-bound, R-bound and Hyperbolic-bound condition.

Even a plethora of efficient solutions do exist in literature but they are mainly tailored for special cases. Like harmonic chain method is strongly recommended for systems, where there are large portion of task periods with multiple relationships. Similarly, if there are still multiple relationships exists in the period belonging to different harmonic chains then the Root Condition is the optimal choice. On the other hand when the period ratio (the ration of maximum and minimum task period values in a task set) is less than 3, the Park-RTA-II condition can be used as an exact test in polynomial time complexity of $O\left(n^3\right)$, otherwise Han-RTA condition is a better condition for systems when the period ration is less than 2.

3 Exact Schedulability Tests

The necessary and sufficient condition for the schedulability in RM can be divided into two categories: scheduling points tests and response time tests [20].

3.1 Scheduling Points Tests

Liu and Layland proved in [16] that the task set is schedulable by RM if the first job of each task can meet its deadline when it is initiated at a critical instant. A critical instant for a periodic task τ_i is the time when τ_i is released simultaneously with request from all higher-priority tasks. Therefore we only need to consider the feasibility of a task set under the worst case phasing. For tasks $\tau_1, \tau_2, \ldots, \tau_i$, the workload (or the cumulative demands) on the processor can be formulated as

$$W_i(t) = C_i + \sum_{j=1}^{i-1} \left\lceil \frac{t}{T_j} \right\rceil C_j. \tag{4}$$

A task τ_i is schedulable in RM if there exists some time $t \in [0, T_i]$ such that

$$\min_{1 \le j \le i} W_i(t) \le t. \tag{5}$$

Define
$$\begin{aligned} L_i(t) &= \quad W_i(t)/t \\ L_i &= \min_{0<t\le T_i} L_i(t) \\ L &= \quad \max_{1\le i\le n} L_i \end{aligned}$$
then the above statement can be expressed as the following theorem [15]:

Theorem 5 *Given periodic tasks* $\tau_1, \tau_2, \ldots, \tau_n$*, (i)* τ_i *can be scheduled for all task phasing using the Rate-Monotonic algorithm if and only if* $L_i \leq 1$*, and (ii) the entire task set can be scheduled for all task phasing using the Rate-Monotonic algorithm if and only if* $L \leq 1$.

Lehoczky et al. first limit the infinite number of scheduling points in the interval $[0, T_i]$ [15], and then show that the function $L_i(t)$ is piecewise monotonically decreasing and the function attains to a local minimum when t is a multiple of one of the periods $T_j (1 \leq j \leq i)$. Hence, to determine whether the task τ_i is schedulable, it only needs to test the points which are the multiples of the periods T_j for $1 \leq j \leq i$. Specifically, let

$$S_i = \{kT_j | j = 1, 2, \ldots, i; \quad k = 1, 2, \cdots, \lfloor T_i/T_j \rfloor\} \tag{6}$$

where k and j are positive integers. The TDA scheduling condition is formally expressed in the following theorem [15].

Theorem 6 *Given periodic tasks* $\tau_1, \tau_2, \ldots, \tau_n$*, (i)* τ_i *can be scheduled for all task phasings using the Rate-Monotonic algorithm if and only if* $L_i = \min_{t \in S_i} W_i t/t \leq 1$*, and (ii) the entire task set can be scheduled for all task phasings using the Rate-Monotonic algorithm if and only if* $L = \max_{\{1 \leq i \leq n\}} L_i \leq 1$.

3.2 Response Time Analysis

Audsley et al. [1] first introduced the exact schedulability test based on RTA. It is an iterative method with a given initial value. Let $R_i^{(n)}$ be the n-th approximation to the exact value of the response time R_i of a task τ_i, in this iteration the $(n + 1)$-th approximation $R_i^{(n+1)}$ can be calculated by the following recursion formula:

$$R_i^{(n+1)} = C_i + \sum_{j=1}^{i-1} \left\lceil R_i^{(n)}/T_j \right\rceil C_j. \tag{7}$$

Except the extended parameters such as worst-case block time, release jitter time, etc., the key study on the RTA-based exact test is how to choose the initial value such that the number of the iterations can be reduced as possible. Eisenbrand and Rothvoß proved that the worst case response time analysis on which the exact tests are based is NP-hard [9]. Therefore the RTA method is not suited to be used for on-line admission control, especially when the number of tasks is large.

3.3 Summary of the Exact Tests

In this section we described two categorises of the exact tests for RM: (i) the scheduling points tests, and (ii) the response time analysis. The main idea of the scheduling points tests is to check the satisfiability of a list of inequalities over a set of finite scheduling points. Generally, researchers focus on reducing the cardinality of the set so that the workload of the tests can be decreased. Interestingly, we notice that although the running time of checking the satisfiability is decreased such approaches need extra time to run the algorithm for obtaining a reduced set of scheduling points. The advantages of investing in reduced set lies in the higher ratio between the adjacent periods. Hence except the original test proposed by Lehoczky et al. [15], the rest of scheduling tests given in Subsection A are recommended to be used when the periods of tasks vary widely. On the contrary, when the differences among periods are small, the Lehoczky et al.'s test is suitable to be utilized. For the response time analysis, the Audsley-RTA is almost replaced by other methods of which the performances vary differently on the task sets with different features in accordance with the literature [19].

4 Rate-Monotonic Optimal Design

Researches on the inexact and exact test are based on the common assumption that all the parameters (task periods and execution times) are fixed. In paper [17], J. Liu et al. define the execution times in intervals instead of fixed points and extend the schedulability test problem to the following new problems:

1. If the task set is not schedulable by RM algorithm, then how to adjust the execution time C_i so that the new task set becomes schedulable.
2. If the task set is schedulable by RM algorithm, then how much freedom do we have to increase the value of the execution time C_i in order to have a better performance while still satisfying schedulability constraints?

The above proposed problems are described as the Rate-Monotonic Optimal Design Problem [17].

C-AP is a generalized form of a question from the sensitivity analysis. The sensitivity analysis is derived from the breakdown utilization presented by Lehoczky et al. [15], where the largest possible scaling values for execution times are calculated and the task set is guaranteed to be schedulable.

There are generally two options to formulate the C-AP. The first is based on the sufficient conditions such as LL-bound and Hyperbolic Bound presented in Sect. 2, and the second is based on the sufficient and necessary conditions such as RTA and HET presented in Sect. 3.

According to the HET condition, the C-AP can be modeled as follows:

$$\max f = \sum_{i=1}^{n} \frac{C_i}{T_i} \tag{8}$$

$$\text{s.t.} \underset{t \in P_{i-1}(T_i)}{\mathrm{V}} \left(\sum_{j=1}^{i} \frac{t}{T_j} C_j - t \leq 0 \right)$$

$$C_i^{\min} \leq C_i \leq C_i^{\max},\ i = 1, 2, \ldots, n$$

Liu et al. [17] proposed the mixed boolean-integer programming (MBP) based method for solving the model (8). In the MBP-based method, J. Liu et al. bring in several integer variables and a big enough positive number to convert the inequalities connected with logic OR into a whole inequality and an equation, as stated in the following theorem.

Theorem 7 *Assume M be a big enough positive number and $h_1, h_2, \ldots, h_m$ be functions. There exist $(y_1, y_2, \ldots, y_m) \in \{0, 1\}^m$ satisfying both $\sum_{i=1}^{m} y_i = m - 1$ and $h_i - y_i M \leq 0 (i = 1, 2, \ldots, m)$ if and only if $(h_1 \leq 0) \vee (h_2 \leq 0) \vee \cdots \vee (h_m \leq 0)$.*

By applying the above theorem, the original model (9) can be equivalently transformed into the following mixed boolean-integer programming problem:

$$\max f = \sum_{i=1}^{n} \frac{C_i}{T_i} \tag{9}$$

$$\text{s.t.} \left\{ \begin{array}{l} \sum_{j=1}^{i} \frac{p_{il_i}}{T_j} C_j - p_{il_i} - y_{il_i} M \leq 0, \\ \quad l_i = 1, 2, \ldots, |P_{i-1}(T_i)| \\ \quad C_i^{\min} \leq C_i \leq C_i^{\max} \qquad\qquad ,\quad \mathrm{i} = 1, 2, \ldots, \mathrm{n} \\ \quad \sum_{k=1}^{P_{i-1}(T_i)} y_{ik} = |P_{i-1}(T_i)| - 1 \\ y_{ik} \in \{0, 1\}, l_i = 1, 2, \ldots, |P_{i-1}(T_i)| \end{array} \right.$$

where y_{ik} is the new involved 0–1 variables, M is the big enough positive number, and p_{il_i} is the l_i-th element of the set $P_{i-1}(T_i)$.

The new model (9) is solved by using the classic branch-and-bound method.

The experimental results illustrate that the MBP-based method is efficient in the case of low task number, while the overhead increases exponentially as the task number raises. Besides the problem of exponential runtime, the MBP-based method has two more disadvantages: (i) when it appears more than one non-integer components in some intermediate point, there has not already been any theoretical guidance on how to choose these non-integer components to construct the branch-problem, and (ii) the method brings new variables of which the number is almost the same as the number of the original variables, so it results in the increase of overhead.

5 Conclusions and Future Work

In this paper, our study showed that for inexact tests, it is recommended to apply a test with lower time complexity in the first place, and proceed to a higher time complexity accordingly when previous test fails. This recommendation is based on the fact that time complexity of a particular test is associated with acceptance ratio-lower is time complexity, lower is the acceptance ratio of the test and vice versa. For exact tests, all the tests are pseudo-polynomial. The scheduling point tests are to check the satisfiability of linear arithmetic formula over some points, and the number of scheduling points should be reduced for lowering the time complexity. RTA is an iterative method, and the number of iterations depends on the initial value therefore it is significant to choose the initial value carefully in order to lower is the complexity as possible.

We also found that Rate-Monotonic Optimal Design Problem has been modelled as an optimization problem extending the fixed execution times in schedulability tests to variables varying in intervals. The modelled optimization problem can handle both the continuous and discrete variables while the traditional schedulability tests are only available for discrete variables and there is a need to investigate the continuous counterpart.

In our future work, we are extended to focus on the following three aspects. First, the survey can be naturally extended to the RM schedulability tests with some relaxed assumptions presented in [2] including the schedulability tests for arbitrary deadlines, dependent tasks with blocking, sporadic tasks, aperiodic tasks and multi-processors and distributed systems. Second, notice that inexact tests may be useful to determine the schedulability of a real-time system with a low processor utilization while exact tests are suitable for the system of high workload. Therefore it is necessary to develop convenient hybrid tests for specific task set in order to reduce the time and space complexities. Third, the decision conditions of the scheduling points tests proposed in [15] can be regarded as linear formulas, and the problem of determining schedulability is equivalent to the problem of Satisfiability Modulo Theories (SMT) [6] over linear arithmetic. It is necessary to study the case of the schedulability testing performed by the SMT algorithms as well as the stat-of-the-art SMT solvers such as MathSat [5], Yices [8]. It would be interesting to make a comparative study on the disjunctive programming algorithm, the OMT solvers and the algorithms for the RM-ODP introduced in this paper by solving more benchmarks as well as the RM-ODP.

Acknowledgements This work is supported by the National Key Research and Development Program of China (No. 2017YFB1400700), the National Natural Science Foundation of China

(No. 61906220) and the National Natural Science Foundation of China (No. 61702570). Meijiao Duan is the corresponding author.

References

1. N.C. Audsley, A. Burns, M. Richardson, K. Tindell, A.J. Wellings, Applying new scheduling theory to static priority pre-emptive scheduling. Softw. Eng. J. **8**(5), 284–292 (1993)
2. S. Baruah, K. Pruhs, Open problems in real-time scheduling. J. Scheduling **13**(6), 577–582 (2010)
3. E. Bini, G.C. Buttazzo, G.M. Buttazzo, Rate monotonic analysis: the hyperbolic bound. IEEE Trans. Comput. **52**(7), 933–942 (2003)
4. D. Chen, A.K. Mok, T.W. Kuo, Utilization bound revisited. IEEE Trans. Comput. **52**(3), 351–361 (2003)
5. A. Cimatti, A. Griggio, B.J. Schaafsma, R. Sebastiani, The mathsat5 smt solver, in *Tools and Algorithms for the Construction and Analysis of Systems* (2013), pp. 93–107
6. L. De Moura, N. Bjørner, Satisfiability modulo theories: introduction and applications. Commun. ACM **54**(9), 69–77 (2011)
7. A. Díaz-Ramírez, P. Mejía-Alvarez, L.E. Leyva-del Foyo, Comprehensive comparison of schedulability tests for uniprocessor rate-monotonic scheduling. J. Appl. Res. Technol. **11**(3), 408–436 (2013)
8. B. Dutertre, *Yices 2.2. Computer Aided Verification*, Lecture Notes in Computer Science (2014), pp. 737–744
9. F. Eisenbrand, T. Rothvoß, Static-priority real-time scheduling: Response time computation is np-hard, in *Proceedings of the Real-Time Systems Symposium* (2008), pp. 397–406
10. C.C. Han, H.Y. Tyan, A better polynomial-time schedulability test for real-time fixed-priority scheduling algorithms, in *Proceedings of the 18th Real-Time Systems Symposium* (1997), pp. 36–45
11. S. Han, M. Park, M. Park, A sufficient condition for rate monotonic schedulability based on response time analysis, in *Proceedings of the International Conference on Computer and Information Technology* (2010), pp. 1751–1757
12. T.W. Kuo, A.K Mok, Load adjustment in adaptive real-time systems, in *Proceedings of the 12th Real-Time Systems Symposium* (1991), pp. 160–170
13. T.W. Kuo, Y.H. Liu, K.J. Lini, Efficient on-line schedulability tests for priority driven real-time systems, in *Proceedings of the 6th Real-Time Technology and Applications Symposium* (2000), pp. 4–13
14. C.G. Lee, L. Sha, A. Peddi, Enhanced utilization bounds for qos management. IEEE Trans. Comput. **53**(2), 187–200 (2004)
15. J. Lehoczky, L. Sha, Y. Ding, The rate monotonic scheduling algorithm: exact characterization and average case behaviour, in *Proceedings of the Real Time Systems Symposium* (1989), pp. 166–171
16. C.L. Liu, J.W. Layland, Scheduling algorithms for multiprogramming in a hard-real-time environment. J. ACM **20**(1), 46–61 (1973)
17. J. Liu, Y. Wang, J. Xing, Study of optimization problems with logic or relationships and its application to real-time system design. J. Softw. **17**(7), 1641–1649 (2006)
18. W.C. Lu, K.J. Lin, H.W. Wei, W.K. Shih, Rate monotonic schedulability tests using period-dependent conditions. Real-Time Syst. **37**(2), 123–138 (2007)
19. W.C. Lu, K.J. Lin, H.W. Wei, W.K. Shih, Efficient exact test for rate-monotonic schedulability using large period-dependent initial values. IEEE Trans. Comput. **57**(5), 648–659 (2008)
20. N. Min-Allah, S.U. Khan, N. Ghani, J. Li, L. Wang, P. Bouvry, A comparative study of rate monotonic schedulability tests. J. Supercomputing **59**(3), 1419–1430 (2012)

21. N. Min-Allah, S.U. Khan, X. Wang, A.Y. Zomaya, Lowest priority first based feasibility analysis of real-time systems. J. Parallel Distributed Comput. **73**(8), 1066–1075 (2013)
22. M. Park, H. Park, An efficient test method for rate monotonic schedulability. IEEE Trans. Comput. **63**(5), 1309–1315 (2014)

The Facility Layout Problem in a Logistics Park Based on Accessibility and Transport Diversity

Lu Qin and Yi Zhao

Abstract A logistic park is a delimited domain having a large space to efficiently organize, manage, and ship goods at a low cost. The facility layout problem in a logistics park is concerned with determining the proper physical organization of a number of related functional areas (FAs). In practical planning, transportation organization is an important consideration, placing the FAs on the plot that best fits it. However, there are few researches based on transport, and the traditional SLP method is increasingly unsuitable for the large-scale development of Logistics parks. Considering this, this study proposes an accessibility-diversity-based method and FA-parcel matching model. Given that the facility layout problem in a logistics park is NP-complete, a heuristic approach of genetic algorithm is presented. A real-world case study is given to test this method. The results show that the both proposed method and solution approach are effective and efficient.

Keywords Facility layout · Accessibility · Transport diversity

1 Introduction

Transportation is a significant consideration in the planning of Logistics Parks. Countries all over the world attach great importance to the site selection of logistics parks. Logistics parks are generally located in areas with convenient transportation, and are close to railway freight stations, expressway entrance and exit, ports, and airports. However, although some logistics parks have convenient transportation conditions, traffic jams at entrances and exits, and safety accidents often occur. One of the reasons is the unreasonable layout of the functional areas (FAs) in the logistics park. Generally, the space of a logistics park is divided into several non-overlapping regions called functional areas. A FA is able to offer one specific logistics service. logistics

L. Qin · Y. Zhao (✉)
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
e-mail: 18120965@bjtu.edu.cn

L. Qin
e-mail: lqin@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_12

parks are commonly configured with five to eight FAs to provide various kinds of service [1]. Some FAs generate a lot of traffic, but some do not. It may cause poor safety and low efficiency if planners do not pay attention to the different transport characteristics of different FAs. Hence, transport analysis is necessary for the layout assignment.

The facility layout problem (FLP) is an arrangement of departments with known dimensions to minimize operating cost and maximize system efficiency [2]. The literature on FLP can be divided into three broad categories. The first category is concerned with algorithms directing at the general FLP. The second category involves the extension of the general FLP that considers additional issues, which arise from real-world applications, such as the optimization of two or three objectives simultaneously and dynamic FLP [3–5]. The third category is concerned with specially structured instances of the problems, like single-row FLP and non-rectangular logistics parks FLP [2].

With the rapid development of logistics parks, the FLP-LP has attracted increasing attention from industry and academia. The method of systematic layout planning (SLP), originally created by Muther Richard, a practical and organized method for rearranging existing or laying out new facilities, was widely applied in FLP-LP [6]. Afterwards SLP was combined with mathematical models and various kinds of algorithms [7, 8]. And it gradually becomes the main method.

Li [9] took Dingzhou railway logistics center as an example, uses SLP method to carry on the layout of it and verifies the feasibility of SLP method. Zhao [10] introduced Markov chain (MC) to improve the traditional SLP, to research and analyze the logistics park function layout planning from five aspects of production, quantity, production route, service department and time. He [11] took a distribution center of as an example, uses SLP method to design its layout. Three sets of layout schemes are obtained by adjusting the actual situation of the DC and the area of each unit, then the scheme is evaluated and selected by the AHP.

However, SLP mainly emphasizes the relevance of work flow, taking the goods flow or closeness between facilities as the basis for layout. The problems of SLP are that there are too much qualitative description or estimation, and its applicability is decreasing for the tendency of logistics parks' increasing scale. Therefore, the SLP is appropriate for planning with small area, such as DC and RDC. Large area makes a logistics park different from a factory or logistics centre when solving the FLP-LP. And the SLP arranges the departments subjectively and qualitatively. There are many companies in logistics parks, having their own business system. The work flow in FAs is not detailed, and the flow of goods between FAs cannot be accurately predicted. Moreover, the method has contradiction with real-world logistics parks because the familiar types of service or goods are often integrated into one FA. It is no longer necessary to load and unload, handle goods between FAs, which increases logistics costs.

It is usually necessary to build an optimal layout scheme by adding constraint conditions and establishing an objective function. Zhao [12] established a multi-objective mathematical model based on the minimum cost of material handling and the maximum degree of adjacency between FAs. Li [13] aimed at minimizing the total transportation costs. Wang [14] established a multi-target facility layout model with the goal of the largest comprehensive relationship, the lowest transportation cost and the lowest land reconstruction cost. However, few researchers consider the transportation conditions.

In this regard, we consider that an accessibility-diversity-based method can effectively be used in handling facilities layout problems in practice. Both accessibility and transport diversity are important factors to be considered in urban planning. The two factors play an important role in road network optimization, land use planning, land price evaluation, location analysis, etc. On the other hand, to our knowledge, accessibility and transport diversity have not been employed to solve the FLP. Considering this gap, an accessibility-diversity-based solution approach for facility layout problem is proposed in this study. We utilize this approach to find the relationships between the FAs and plots in the logistics parks.

The paper is further organized as follows. Section 2 describes the accessibility-diversity-based method. Section 3 utilizes heuristics algorithms to solve the proposed model. A real-world case study is presented in Sect. 4. Finally, conclusions and future research directions are given in Sect. 5.

2 Methodology

2.1 FAs Transport Requirement Measurement

2.1.1 Accessibility Requirement Measurement

Accessibility is the resistance to overcome spatial obstructions of different regions by using transportation system [15], and it is the degree of difficulty to reach various types of activity areas from a certain point [16]. Therefore, the accessibility requirement of the FAs refers to the expectations of goods in the area for low-cost, fast and efficient transporting to transport hubs and highways. The traffic requirements of goods form the FAs' traffic requirements characteristics which are determined by three factors, namely FAs' freight volume, shipping frequency and drayage (short distance movement of freight) timeliness requirements (DTR). The three factors influence the FAs accessibility simultaneously. The freight volume refers to the total amount of goods shipped out of a FA per unit time. It does not consider the consumption of goods in the FA, and is calculated separately according to different transport modes. Shipping frequency is a supplement and improvement to the factor freight volume. It refers to the number of times the goods delivered in a unit time. Frequent shipments of a FA mean that it has strong needs for accessibility. The DTR refers to the time required for the goods to depart from the FA to the surrounding transport hubs or directly to the destination. In order to quantitatively evaluate this degree of requirement, it is divided into three levels: the first level is to serve the city logistics, the second level is to deliver the goods to the railway freight station, port or airport with features of fixed time, fixed location, and fixed route. The third level is no strong demand to drayage timeliness because the total transport time is too long to bother about the drayage time.

Let us consider a logistics park, which has M FAs, where M denotes the number of the FAs. N denotes the number of transportation modes of FAs. AR_{ij} denotes accessibility requirement of FA i to transport hub or highways j, where $(i \in 1, 2, \ldots, M, j \in 1, 2, \ldots, N)$. Moreover, Q_{ij} denotes the freight volume of FA i to j; f_i denotes the shipping frequency of i and T_i denotes the average inventory cycle of i; θ_i denotes drayage timeliness requirement index which the first, second and third level are let to be 0.8–1.0, 0.5–0.8 and 0.1–0.5. We present an equation for the accessibility requirement of FA i to j as follows:

$$\mathrm{AR}_{ij} = \theta_i \cdot Q_{ij} \cdot f_i = \theta_i \cdot Q_{ij} / T_i \tag{1}$$

2.1.2 Transport Diversity Requirement Measurement

The requirement for transport diversity of FAs refers to the expectations of goods in the FAs for free selection of multiple transportation modes to their destinations, and is the degree of dependence of goods on diversified modes of transportation. When the goods in the FAs have a strong demand for transportation by road, rail and air multiply, the transport diversity requirement of the FA is strong; However, when all the goods in the FA are carried simply by road, the transport diversity needs are quite weak. Based on the logic of Herfindahl-Hirschman Index (HHI) which is a comprehensive index for measuring industry concentration, we present an equation for transport diversity requirement as follows:

$$\mathrm{DR}_i = 1 / \sum_{j=1}^{N} \left(\frac{Q_{ij}}{Q_i} \right)^2 \tag{2}$$

where DR_i represents the transport diversity requirement of FA i, and Q_i is denotes the freight volume of FA i. Equation (2) represents the freight structure concentration of the FA. The simpler the freight transportation mode of a FA is, the higher freight structure concentration is, the weaker the requirement for transport diversity of the FA gets; otherwise, the requirement is stronger.

2.2 Parcels Traffic Conditions Measurement

2.2.1 Accessibility Measurement

Logistics parks are commonly divided into multiple parcels by road networks and natural obstructions. Parcels accessibility refers to the convenience which parcels in the logistics park can reach surrounding transport hubs (airport, port, railway freight station, etc.) and highways.

With regard to transportation hubs which can be abstracted, simplified as a point, the parcel accessibility to those points is evaluated by classical models, such as spatial separation measure, cumulative opportunities, spatial interaction measure, utility measure, etc. The spatial interaction measure is defined as the potential for interaction between two points [16], considering accessibility is not only related to spatial separation of the two points, but also the scale of terminal points. We modify this model by using the effective transport time as the impedance function, and taking the appeal of these three modes of transportation to the park is their freight volume which are 10.1, 30.0, and 12.7 million tons per year, the detailed approach is as follows:

- Step 1. divide the logistics park's land to several parcels. Like the division of traffic zones, the division should consider natural and artificial obstructions as dividing lines which are still in coordination with the road network (existing or in plan); the number of parcels should be appropriate, between 8 and 15 subject to the size of the logistics park; the size of each parcel should be moderate and not be larger than the smallest FA to ensure that the parcels will not be separated furthest.
- Step 2. measure the shortest transport time from the center of each parcel to each transport hub. In most cases the goods will take drayage by road so that the shortest transport time is measured based on the actual distance of the shortest path and the average speed of vehicles, not considering traffic congestion and other kinds of waiting time.
- Step 3. calculate the accessibility of each parcel to each transport hub. Based on the spatial intersection measure, a formulation is presented as follows:

$$A_{kj} = q_j / T_{kj}^{\alpha} \tag{3}$$

where A_{kj} represents accessibility of the parcel k to transport hub j; q_j denotes annual freight volume of j; T_{kj} denotes the shortest transport time from the parcel k to j, where $(k \in 1, 2, \ldots, K)$; α is a parameter of impedance influence degree, whose value is normally 2 [9]; K denotes the number of parcels.

However, with regard to roads which cannot be simplified as points, the spatial interaction measure is not applicable. This study uses Space Syntax theory to calculate the road accessibility of parcels. Space syntax is a set of techniques for analyzing spatial layouts and human activity patterns in buildings and urban areas, often used to study the structure of road network, land use intensity, and spatial distribution of land [17]. Connectivity, topological depth and integration are indexes of the characteristics of the road network structure [18]. We use Global Integration Value (GInteg) to measure the accessibility of the parcel to roads. GInteg indicates the closeness degree of the road (which is the closest to the FA) and the other roads in the logistics park [19]. The higher the value of GInteg is, the higher the road accessibility of the parcel is. The following formula is used to determine the GInteg:

$$A_{kj} = \mathrm{GInteg}_k = \frac{n\left[\log_2\left(\frac{(n+2)}{3} - 1\right)\right] + 1}{(n-1)(MD_i - 1)} \tag{4}$$

where GInteg_k denotes the Global Integration Value of the parcel k; MD_k denotes the average topological depth value of the parcel k, representing the average value of the shortest distance from random axes in the boundary of the study to the axis which is the closest to the parcel k; n denotes the number of axes in the study boundary.

2.2.2 Transport Diversity Measurement

Transport diversity of a parcel represents the degree of freedom to select multiple modes of transport. The accessibility of all transportation modes of the parcel determines the potential to choose diversified transportation modes. This study uses the weighted average of accessibility of all transportation modes as a comprehensive value to measure the transport diversity of the parcel. We use the entropy measure for the influence of the various accessibility is unclear. The entropy measure is a method to determine the weight of an index. Let D_k denotes the transport diversity of the parcel k; A_{kj} denotes accessibility of transportation mode j of the parcel k; H_j denotes the entropy value of j; w_j denotes the weight of j. The calculation process is given as follows:

$$p\left(A_{kj}\right) = \frac{A_{kj}}{\sum_{k=1}^{K} A_{kj}} \tag{5}$$

$$H_j = -\frac{1}{\ln N} \sum_{k=1}^{K} p\left(A_{kj}\right) \ln p\left(A_{kj}\right) \tag{6}$$

$$w_j = \frac{1 - H_j}{N - \sum_{j=1}^{N} H_j} \tag{7}$$

$$D_k = \sum_{j=1}^{N} w_j A_{kj} \tag{8}$$

Equation (5) is used to normalize various accessibility values; Eq. (6) calculate the entropy value of each index j; Eq. (7) convert the entropy value into weight. Equation (8) is used to calculate the weighted average value, namely the transport diversity of the parcel k.

2.3 Accessibility-Diversity-Based Model

The problem can be described as: there are M planned FAs and K parcels ($K \geq M$) in the logistics park. It is required to reasonably arrange the FAs to the parcels, meeting the area requirements of the FAs, and maximally ensuring that the parcels match with the FAs. It is a multi-objective optimization problem because both accessibility and transport diversity matching degree should be optimal. There are several assumptions as follows: (1) We assume that each FA is a traffic zone, and the freight flow of the cell is generated by the zone's centroid; (2) Parcels are the smallest units that cannot be separated, so that one parcel cannot be selected by two or more FAs. Once all FAs' accessibility, transport diversity requirement and all parcels' accessibility, transport diversity is measured, the following model is used to determine the layout of FAs:

$$\max U = \sum_{i=1}^{M}\sum_{k=1}^{K}\lambda_{ik}\left(p\cdot\frac{\sum_{j=1}^{N_i} AR_{ij}\cdot A_{kj}}{N_i} + q\cdot DR_i\cdot D_k\right) \tag{9}$$

$$\text{s.t.}\ \sum_{i=1}^{M}\lambda_{ik} = 1 \tag{10}$$

$$\sum_{k=1}^{K}\lambda_{ik} \geq 1 \tag{11}$$

$$\sum_{k=1}^{K}\lambda_{ik}\cdot \mathrm{SP}_k \geq \mathrm{SF}_i \tag{12}$$

$$p + q = 1 \tag{13}$$

where U is the global matching value; p and q is the weight of the accessibility and transport diversity, used to converting this multi-objective problem into a single-objective problem; SF_i and SP_k denotes the area of the FA i and parcel k. λ_{ik} is a 0–1 variable, indicating the parcel k is whether selected by the FA i or not. The objective function of the model is to maximize the matching value of global accessibility and transport diversity between parcels and FAs and these values should be normalized. Constraint (10) ensures that each parcel is only selected by one FA; Constraint (11) ensures that every FA should select one parcel at least; Constraint (12) indicate that the total area of the FA allocated to parcels is not larger than that of the parcel; Constraint (13) shows the weight of the accessibility (p) and transport diversity (q).

3 Solutions Algorithm

Since the allocation problem belongs to the class of NP-complete problems, it is a challenge for an exact algorithm to find the optimal solution for a large problem within a reasonable time. Therefore, heuristic algorithms are employed to solve it. Lin [20] applied genetic algorithm (GA) earlier to solve the FLP. Because GA has the best searchability globally, it can effectively avoid falling into a local minimum, so it is widely used in the subsequent time period. Compare to Simulated Annealing (SA) and Hill Climbing, GA can efficiently search for the globally optimal solution. So the article selects GA to solve the problem. The details of the algorithm are provided by the following steps:

Step 1. Genetic encoding

A chromosome represents a solution to the problem in GA. Typically the chromosome is a string of genes coded by binary values or Arabic numbers. To handle the constraints in the model, the chromosome is designed to be composed of K genes where K is the number of the parcels. Each genes represents the FA's number marked by us and the genes are ranged according to the number of parcels by Arabic number in an increasing order from 1 to K. Given a logistics park, three FAs placed in four parcels. The solution indicated by chromosome "2313" means that FA 2 is placed in Parcel 1, FA 1 is placed in Parcel 3, and FA 3 is placed in Parcel 2, 4.

Step 2. Population initialization

80 chromosomes are randomly generated as the initial population, representing a set of initial candidates.

Step 3. Fitness function formulation

A fitness function is to identify the efficiency of a solution to achieve. It generally requires non-negative, the objective function (9) is a non-negative function and looking for the maximum value. Therefore we can directly use the objective function as the fitness function. The fitness function is developed as follows:

$$f = p \cdot \frac{\sum_{j=1}^{N_i} \mathrm{AR}_{ij} \cdot A_{kj}}{N_i} + q \cdot \mathrm{DR}_i \cdot D_k \tag{14}$$

Step 4. Reference set initialization

A reference set is a set of good and diverse solutions, selected from set obtained from Step. 2. The good solutions can be selected based on the fitness function.

Step 5. Subset generation

Generally, GA chooses two elements from the reference set to create new solutions. With this mechanism, the subsets can be generated.

Step 6. Crossover and mutation

Usually, a high crossover rate may ruin good chromosome structures, and a low mutation rate may be useless to the diversity. The crossover generally ranges from 0.25 to 0.95, let it be 0.7 in this study; the mutation generally ranges from 0.0001 to 0.1, let it be 0.01 in this study.

Due to the use of real number coding, there will be duplicate gene problems. This article uses the allele method to solve the problem of duplicate genes in crossover operations. When performing the crossover operation, after judging the occurrence of the duplicate gene, replace the occurrence of the duplicate gene to the same locus in the two individuals, and then perform the crossover operation, so as to avoid the occurrence of duplicate genes in the newly generated offspring. Regardless of the location of the intersection point, the intersection operation can be performed at any intersection point, and the allele method has no effect on the fitness value of the individual, which solves the problem of duplicate genes.

Step 7. Reference set updating

A static updating approach is employed in this step. From the current reference set, several good solutions and diverse solutions are chosen, and then new chromosomes are acquired through crossover and mutation to obtain a new reference set. If two consecutive reference sets are the same or the iterations reaches 200; otherwise, the procedure returns to Step 5.

4 Case Study

In order to explore the practicality of the method, this study chooses a real-world case of logistics park which is located in an industrial park in Chongqing, China, with a total area of 246 ha. There is a railway freight station on the north and an inland river port on the east. Six FAs are planned in this logistics park which are Highway Loading Area (HLA), MRO Storage Area (MROSA), Large-quantity Cargoes Storage Area (LQCA), Goods Vehicles Storage Area (GVSA), Bonded Logistics Center (BLC), and Container Area (CA). The goods characters of storage and transport of the six FAs are quite different (Table 1). The normalized values of accessibility and transport diversity requirement (Table 2) are obtained from the method in Sect. 2.1. In general, HLA and CA have strong demand to railway, port accessibility; HLA has strong demand to road accessibility. These FAs have high freight volumes and short storage cycle. However, LQCA, GVSA and BLC have weak demand to accessibility for they have low freight volumes, long storage cycle and low drayage timeliness requirement. In regard of transport diversity, CA, GVSA, LQCA and BLC have

Table 1 Basic information of the FA

No.	FA	Freight volume (10,000 tons)	Area (ha)	Average storage cycle (day)	DTR	Transport proportion		
						Road (%)	Railway (%)	Waterway (%)
1	HLA	545	30	4	0.25	80	13	8
2	MROSA	479	32	7	0.14	78	15	8
3	LQCA	206	40	6	0.17	55	28	18
4	GVSA	200	53	18	0.06	78	8	15
5	BLC	250	40	40	0.03	35	45	20
6	CA	312	43	5	0.20	40	30	30

Note The freight volume and area are obtained from investigation and planning; the average storage cycle is estimated by the cargo flow characteristics and some similar FAs in other logistics parks; the DTR is set according to the proportion of each transportation mode

Table 2 FAs accessibility and transport diversity value

No.	FA	Rail AR	Port AR	Road AR	Transport Diversity
1	HLA	2.2	1.9	3.4	1.51
2	MROSA	1.3	0.9	1.6	1.59
3	LQCA	0.7	0.7	0.3	2.45
4	GVSA	0.0	0.1	0.1	1.59
5	BLC	0.1	0.1	0.0	2.74
6	CA	1.7	2.4	0.5	2.94

strong demand; HLA, MROSA and GVSA have weak demand because of simple transport structure.

Confirm and simplify the boundaries of the logistics park, and divide the land into 13 parcels according to the traffic network and the restrictions of the high-voltage corridor which becomes the black parcel shown in Fig. 1. Parcels that are too small (smaller than 5 ha) will be alternative but not be counted. The accessibility is measured in tree transportation.

Modes: rail, port, and highway. The freight volume of these three modes to the logistics park is that their freight volumes are respectively 10.1, 30.0, and 12.7 million tons.

To measure the road accessibility, first, use AutoCAD to build the road network structure of the park, and use Depth Map to calculate the global integration of the parcels. Port and rail accessibility values are obtained from the method in Sect. 2.2. The results are shown in Table 3 and Figs. 2, 3, 4 and 5. The darker the plot, the larger the transportation accessibility or diversity values, and the better the traffic conditions. In general, the traffic conditions on the north side are better than the south side, and the east side is better than the west side.

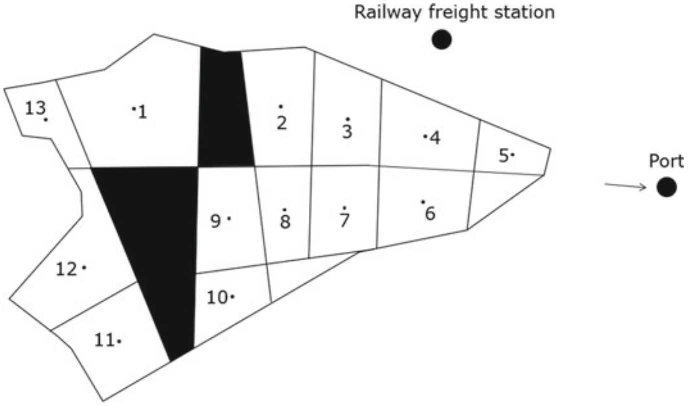
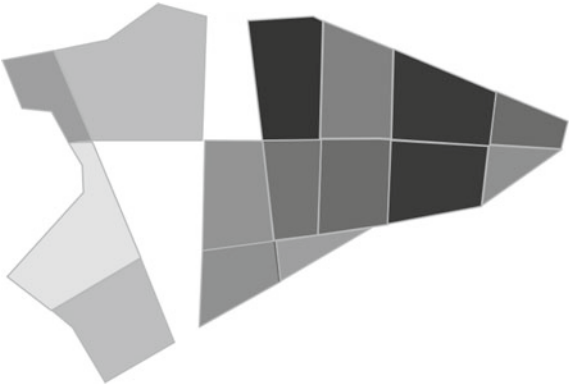


Fig. 1 Logistics park parcels diagram

Table 3 Parcels accessibility of three transportation modes and transport diversity values

No.	Area (ha)	Rail	Port	Road	Transport diversity
1	29.5	87.9	30.4	1.32	7.7
2	22.2	46.9	33.2	1.85	7.0
3	20.0	43.4	33.8	2.77	7.9
4	20.2	96.1	41.2	1.82	9.6
5	10.0	200.7	45.4	1.69	14.3
6	18.0	81.7	39.7	1.85	8.8
7	16.0	44.4	34.8	1.87	6.9
8	15.9	46.4	34.3	2.02	7.1
9	21.1	37.0	31.5	1.95	6.6
10	13.7	30.5	31.7	1.60	5.7
11	19.3	21.5	27.7	1.20	4.5
12	30.8	29.3	24.8	1.78	5.6

Fig. 2 Road accessibility diagram



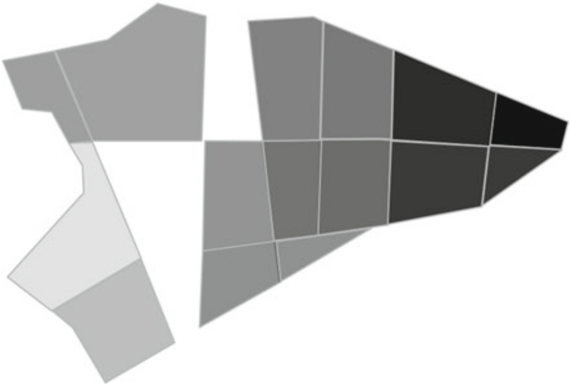


Fig. 3 Port accessibility diagram

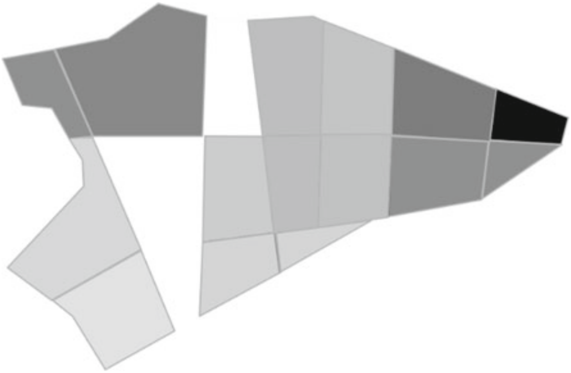


Fig. 4 Rail accessibility diagram

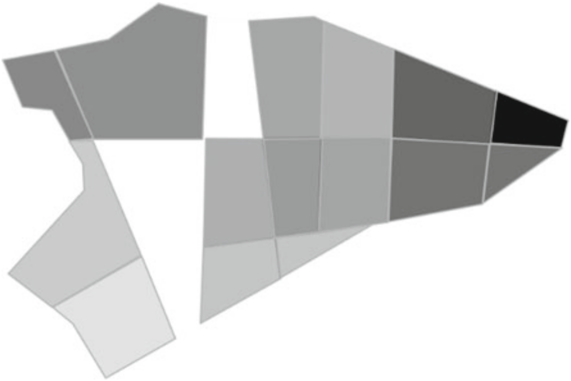


Fig. 5 Transport diversity diagram

In this case, traffic accessibility and diversity are considered equally important, so p = q = 0.5. Use Matlab2014 to run the GA for ten times. The population number is set to 80 and the number of iterations is 200. The best value is 30.65 operated within 10 s. The operation result is shown in Fig. 6. The layout results obtained from the chromosomes are shown in Fig. 7. In the study of the facility layout of logistics parks, there are no uniform and objective indexes for evaluating the layout results. Therefore, we cannot compare the proposed methods and results with other methods and results. However, comparing with the manual operation, the FA-parcel matching model improves the layout efficiency and meanwhile reducing the subjective uncertainties of the SLP method.

In the future, the research on the accessibility and diversity can be more detailed, and given weights by experts not just calculating freight volume. And the algorithm can be combined with GA and others such as SA, Ant Colony Optimization (ACO) to make the solution more efficient.

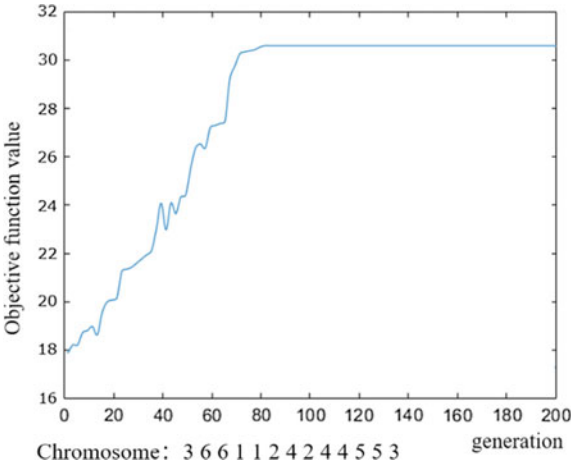


Fig. 6 Model operation results

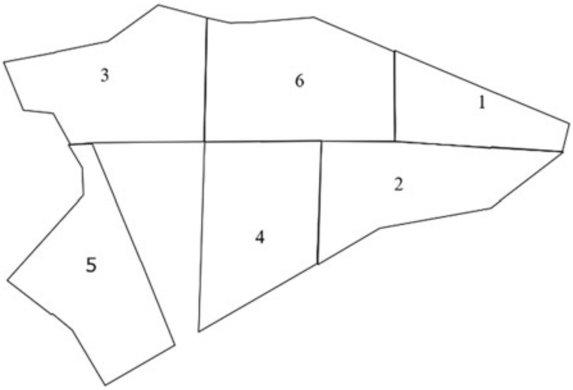


Fig. 7 Layout results

5 Conclusions

Facility layout is quite significant for logistic parks planning due to its direct effects on efficiency and convenience. However, FA layout is a difficult activity since both qualitative and quantitative factors affect the location decision. Transportation is a key consideration among them. However, there are few studies on the layout problems based on transportation. It is a new direction to utilize the accessibility and transport diversity in FLP, quantifying the influence of transportation to FA layout. This study defines the accessibility and transport diversity of FAs and parcels, and uses several approaches to measure them. We propose a FA-parcel matching model which will be solved by GA. In order to verify the applicability of the method, a practical case is applied, and the results show that the method can effectively solve the layout problem.

References

1. C. Liang, Y. Yang, X.C. Wang, *Logistics Park Planning* (China Fortune Press, Beijing, 2013)
2. Y. Chen, et al., The facility layout problem in non-rectangular logistics parks with split lines. Expert Syst. Appl. **42**(21), 7768–7780 (2015)
3. M.J. Rosenblatt, The dynamics of plant layout. Manage. Sci. **32**(1), 76–86 (1986)
4. H. Pourvaziri, B. Naderi, A hybrid multi-population genetic algorithm for the dynamic facility layout problem. Appl. Soft Comput. **24**, 457–469 (2014)
5. B. Ulutas, A.A. Islier, Dynamic facility layout problem in footwear industry. J. Manuf. Syst. **36**, 55–61 (2015)
6. R. Muther, *Systematic Layout Planning* (Industrial Education Institute, Boston, MA, 1961)
7. F. Feng, L. Jing, L. Yang, Layout method for the functional area of railway logistics center based on the improved systematic layout planning. China Railway Sci. **33**(2), 121–128 (2012)
8. Y. Sun, C. Ma, W. Zhang, Research on layout problem in logistics park based on shortest path and multi-population genetic algorithm. Logistics Sci-Tech **38**(2), 86–91 (2015)
9. J. Li, Research on the plane layout planning of railway logistics center based on the system layout planning-ingzhou railway logistics center as an example. Logistics Eng. Manage. **40**(3), 74–76 (2018)
10. J. Zhao, N. Lye, Layout of logistics park based on improved SLP method. J. Chang'an Univ. Natural Sci. Edn. **40**(3), 100–108 (2020)
11. Y. He, M. Wu, Layout planning and design of an E-commerce logistics distribution center. Logistics Sci-Tech **41**(8), 52–55 (2018)
12. B. Zhao, Research on layout planning of agricultural products cold chain logistics park. Logistics Sci-Tech **42**(12), 140–143 (2019)
13. J. Li, Modeling and realization of the spatial layout of Beijing-Tianjin-Hebei iron and steel logistics park. J. Beijing Univ. Technol. Soc. Sci. Edn. **17**(3), 1–7 (2017)
14. J. Wang, A study of the functional area layout planning of logistics park considering railway dividing line. Industrial Eng. J. **22**(5), 102–108 (2019)
15. J.M. Morris, P.I. Dumble, M.R. Wigan, Accessibility indicators for transport planning. Transp. Res. Part A: General **13**(2), 91–109 (1979)
16. W. Hansen, How accessibility shapes land use. J. Am. Institute Planners **25**(2), 73–76 (1959)
17. C. Alalouch, et al, The impact of space syntax spatial attributes on urban land use in Muscat: Implications for urban sustainability. Sustain. Cities Soc. **46**, 101–115 (2019)

18. H. Wu, A. Tsai, H. Wu, A hybrid multi-criteria decision analysis approach for environmental performance evaluation: an example of the TFT-LCD manufacturers in Taiwan. Environ. Eng. Manage. J. **18**, 597–616 (2019)
19. Z. Qi, et al., Evaluation on the accessibility of the scenic spots in Wuhan based on the spatial syntax. Econ. Geogr. **35**(8), 200–208 (2015)
20. L.K. Nozick, M.A. Turnquist, Inventory, transportation, service quality and the location of distribution centers. Eur. J. Oper. Res. **129**(2), 362–371 (2001)

The Study of the Mode and Operation Mechanism of Government Procurement Management in China

Bing Hou, Ying Chai, and Xiao Xiao

Abstract The historical stage of government procurement in China has been systematically concluded in this essay as well as the type and operation mechanism of government procurement management mode. The problems of government procurement management mode are discussed and there is the analysis of the causes of problem. Based on the current situation of government procurement in China and the interests of the public, the AHP method is used to construct a stakeholder-based government procurement performance evaluation model, which includes 3 levels such as target level, guideline Level and assessment level. It also includes 6 parts which are economy, personnel quality, technical management guarantee, social responsibility, stakeholder evaluation, effectiveness evaluation; and 17 specific indicators. The empirical research has been done to support the analysis.

Keywords Government procurement · Management mode · Operation mechanism · Stakeholders · Performance evaluation · AHP

Government procurement refers to the purchase of goods, works and services by governments at all levels from domestic and foreign markets for government departments or affiliated groups in domestic and foreign markets by means of public tendering and fair competition, in a legal manner, methods and procedures, under the supervision of the government, in order to carry out daily government affairs activities or provide services to the public [1]. As an advanced public financial expenditure management system, government procurement has the more than 200 years history

B. Hou
Office of State Assets Management, Beijing Jiaotong University, Beijing, China
e-mail: binghou@bjtu.edu.cn

Y. Chai (✉)
Office of Academic Affairs, Beijing Jiaotong University, Beijing, China
e-mail: ychai@bjtu.edu.cn

X. Xiao
Institutional Banking Department, China Construction Bank, Beijing, China
e-mail: xiaoxiao.zh@ccb.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_13

in foreign countries, and gradually established a more perfect legal system of government procurement. Although the construction of the government procurement system in China started quite late, after 20 years of development and reform, the Government Procurement Law, the Bidding Law of the People's Republic of China and other government procurement core legal system have been established successfully which basically formed a combination of 'unified supervision, separation of management, centralized procurement and decentralized procurement government procurement system. With the development of China's economy and the extensive use of Internet technology in the field of government procurement, the rapid development of government procurement and the scale of it is gradually expanding.

1 The Historical Stage of Government Procurement in China

The construction and selection of the government procurement management model in China took the pilot reform in Shanghai as the starting point, and has experienced four historical stages: pilot exploration(July 1996–July 1998), establishment and promotion (July 1998–January 2003), rapid development (January 2003–September 2015) and reform perfection (September 2015–present).

In the first stage, Shanghai groped the way across the river through the initial period of management mode construction. The embryonic form of decentralized management mode which contains the government procurement administrative institutions (government procurement management office) and the centralized procurement institutions (government procurement centre) has been set up during the first period [2].

In the second phase, 29 provinces, autonomous regions, municipalities and cities began to form and promote government procurement. However, some provinces implemented by using a different method from Shanghai which was moving the centralized procurement institutions under the financial sectors. Or giving all the responsibilities of supervision and implementation to the procurement centre to promote a more integrated and centralized management mode.

In the third stage, the Government Procurement Law of the People's Republic of China (here in after referred to as the Government Procurement Law) was promulgated and implemented in January 2003, which strongly promoted the rapid development of government procurement. Most provinces implemented a decentralized management mode of the separation of institutions and management and an independent operation in accordance with the requirements of the Government Procurement Law.

As for the fourth stage, the target was to build a service-oriented government which can provide 'one-stop civil services, promote the integration of government procurement resources in an orderly manner, unify the rules of government procurement transactions, improve the operation mechanism of government procurement, the government procurement organizations and supervision system, and strengthen the security of implementation. On 15th Sep, 2015, the Ministry of Finance issued the Notice on The Implementation of the Work Programme on the Integration and Establishment of a Unified Public Resources Trading Platform no. 163 of the Treasury(2015), clarified the related work such as the integration of government procurement would be brought into the public resources trading platform. It also represented that the construction of the government procurement management mode has entered the improvement period of "full disclosure online and no transactions offline".

2 The Type of Current Management Mode of Government Procurement in China

Under China's current government procurement system, different business procurement is divided into different management departments. The National Development and Reform Commission is responsible for the key construction project; the Ministry of Construction is taking the responsibility for engineering project; the Health Department is in charge of all pharmaceutical projects; other projects included in the centralized procurement list are managed by the government procurement regulatory institutions in the Department of Finance to organize the government centralized procurement institutions or social agencies who are responsible for procurement. Since the Government Procurement Law does not make clear on the establishment of government procurement institutions in China, which only requires that the establishment should be based on local demand. Therefore, the setup for government procurement institutions vary from place to place. During the pilot start-up period, Shanghai established a decentralized management mode which is setting up the government procurement authority (hereinafter referred to as the Procurement Office) and the centralized procurement institutions (the government procurement centre). Subsequently, there were 29 provinces, autonomous regions, municipalities and cities started to set up and promote the government procurement.

At this point, the regulatory authority of the financial department has the responsibility of business supervision and guidance to the centralized procurement institutions and social agencies that carry out government procurement. It also dealt with the centralized management of the social agencies within the area. According to the requirements of the Government Procurement Law, the executive director of the centralized procurement agency that implements 'management separation' is different from place to place, and there are currently four forms of setting up: First, establishing an independent centralized procurement institution that do not have a affiliation with the administrations, such as the Shanghai Municipal Government Procurement Centre which was led by the Shanghai Municipal Government Procurement Committee. Secondly, setting up the centralized procurement agencies managed by relevant units, such as the Beijing Municipal Government Procurement Centre is managed by the Beijing Municipal People's Government State-owned Assets Supervision and Administration Commission; Heilongjiang, Liaoning, Inner Mongolia Autonomous Region and other centralized procurement institutions are monitored by their provincial government. But Jilin Province is managed by the provincial government affairs bureau, provincial government office, provincial government services and digital construction authority [3]. Next, combining the government centralized procurement work with other state-owned assets trading activities into a unified platform to operate. For example, in Shaoxing City Zhejiang Province, the land transactions, construction project transactions, property rights transactions, government procurement and other activities were organized together to practice on the Shaoxing bidding centre platform. At last, without setting up the centralized procurement institutions, the entrusted social agencies are taking over the centralized procurement tasks.

This four setting-ups, from a longer perspective, are not beneficial to the self-construction of centralized procurement institutions. They also made it difficult to motivate initiatives for their own exploration and development. On the one hand, the authority of centralized procurement is hard to be formed which leads to the lack of supervision and restriction of the buyer's tendentious behaviour. During the organization of procurement activities, if the procurement agency does not have the power of supervision or restriction, the purchaser's behaviour disposition or unreasonable requirements will be put on the cover of legalization by the standardized process, which will lose the fairness of government procurement. The value and significance of the existence of centralized procurement institutions will be vanished. On the other hand, centralized procurement agencies in different cities do not have a unified administrative department which makes it not that easy for sharing and exchanging all kinds of resources and business experience. Also, the joint force and advantages of industry cannot be neither combined nor developed. Furthermore, it will not contribute to the practice and extension of the government procurement's social, economic policy functions. In a word, there are two types of government procurement management modes which are the most representative: 'strictly decentralized' and 'regional decentralized'.

3 The Operation Mechanism of Government Procurement in China

During the exploration period of the reform of government procurement system in 1996, the management and executive institution system of government procurement was initially established, as well as the operation mechanism. The financial department, discipline inspection and supervision department and audit department were formed to work and supervise together [4].

3.1 Planning Mechanism

In government procurement organizations, managers at each level should be involved in planning activities in the organization's procurement activities. The government procurement budget is the plan of both the government annual procurement project and the use of funds in the administrative institution. It is an integral part of the department budget. The preparation of government procurement budget is the premise and foundation of the implementation of government procurement which is also an effective measure to regulate the procurement behaviour of purchasers [5]. It is not only the content of public finance but also the basis of carrying out government procurement activities. At the same time, it can help with the reduction of unplanned and repetitive procurement by providing the institutional guarantee. The government procurement plan is the specific program of the government procurement budget, including the specific project of government procurement, the organization form of the procurement project, the purchase method, the composition of the purchase fund, the method of payment, etc. The government procurement plan has a significant effect on strengthening the management of government procurement, regulating the behaviour of participating stakeholders in government procurement and completing the procurement task. It is also the requirement to realize the goal of government procurement to specify the government procurement goal and task.

3.2 Organizational Mechanism

As an element of the operation of the government procurement management system, the main function of organizational mechanism is to divide and determine the functions and tasks, coordinate and restrain the behaviour of the various functional departments of government procurement institution, regulatory body and individual employees according to certain principles and appropriate forms. In accordance with the provisions of the Government Procurement Law, China's financial departments at all levels are the supervision and management bodies of government procurement. The centralized government procurement agencies set up on demand are the executive bodies of government procurement. As a regulatory body, the financial department has the function of formulating and implementing government procurement policy. It has the function of budget approval for government procurement and managing the state's fiscal revenues and expenditures [6]. At the same time, the department of finance is also in charge of the purchaser's procurement plan for approval, the use of its funds for the final accounts audit, all kinds of government procurement complaints, plan approval, auditing the plan and the implementation of the plan which leads to the financial sector's getting powerless and its policy-making getting less professional.

3.3 Coordination Mechanism

Government procurement is a complex transaction process involving multiple stakeholders and the coordination runs through the whole process of government procurement. At present, China's government procurement has not got a special government procurement coordination mechanism yet. All levels are advised to refer to the Government Procurement Law and Government Procurement Law Implementation Regulations once there are relevant coordination problems during the organization of government procurement activities. However, the scope of responsibility does not include goods and bidding services procured by government. This kind of drawback of the system, resulting in the existing responsibility of government procurement is not able to effectively solve the problem through mutual constraint on coordination mechanism.

3.4 Control Mechanism

The control mechanism of government procurement is mainly embodied in the supervision of the regulatory body and the executive body. Briefly, it is to monitor the process of administration and management and operation results of government procurement in order to prevent and correct the mistakes and deviations during the procedure, and to ensure the effective operation of government procurement administration. The main body of supervision in China is the National People's Congress, the audit department, the discipline inspection department and the supervision department. In the process of implementing and controlling government procurement activities, the NPC is responsible for the consideration and supervision of the government procurement budget, the audit department is responsible for the full supervision of the implementation of the budget and other income or expenditure. The discipline inspection and supervision department is responsible for the inspection and supervision of the compliance of the relevant administrative bureaus in government procurement activities and the supervision of the law and regulations.

4 The Analysis of Problems and Causes of the Government Procurement Operation Mechanism

According to the operation mechanism of the current management mode of government procurement in China, no matter using the strict decentralized management mode or the implementation of regional decentralized management mode, there are still some common problems which are budget irregularities, unclear responsibilities, lack of coordination mechanism and the shortage of supervision.

4.1 The System Is not Perfect. There Is not a Standard for the Budget and Plan

The government procurement budget and plan are the first guarantee for government procurement system. Whether there is sufficient market information and whether it can obtain evaluation feedback on time are the objective constraint factor of the scientific design for the government procurement plan. Under the background of the relatively decentralized management mode, it is difficult to form the effective data support, which causes the institutional obstacle in the design stage. In addition, the parallel setting of institutions has led to the lack of adequate technical support among various institutions, especially the lack of staff ingenuity with a certain professional level in grassroots units; meanwhile, both inconsistent standard of preparation and the irregular preparation have further affect the quality of procurement budget and planning. It can be seen that only in the budgeting aspect, the preparation standards, preparation content, staffing and procurement needs are unified that it will enable limited technical personnel to maximize the effectiveness of the basis of adequate and perfect the information reserves to make a scientific procurement plan.

4.2 The Unequal Power and Responsibility and Unbalanced Power Operation

The unequal power and responsibility are one of the important factors restricting the development of government procurement to become standardized and scientific. The concentration of power is the last choice in the era of planned economy. In the period of relatively scarce resources, it is necessary to rely on powerful dispatch to achieve government management goals. With the development of market economy and the transformation of government function from management to service, the allocation of power gradually returns to the principle of reciprocity of power and responsibility. Under the background of this era, the purpose of government procurement has also experienced a change in the concept of guarantee to management then to service. In this process, the research on government procurement has changed from the initial emphasis on the power of the purchaser to the maintenance of the power of the purchaser then to the ultimate checks and balances the power of the purchaser, which is in line with the development of government procurement. At the same time, the establishment of the agency not only helps with improving the procurement efficiency and standardizing the procurement process, but also improving the imbalance of rights and responsibilities in the past to a certain extent. However, it still fails to achieve a reasonable distribution of rights and responsibilities. In the actual operation process, there are still excessive intervention of purchaser's power greater than responsibility and agencies' responsibility greater than power. The current 'island management' leads to the lack of systematic and effective regulatory mechanism and accountability mechanism, which objectively increases the risk of abuse of power and subjectively increases the doubt on the rationality of decision-making. Supplier's unreasonable complaints take up a considerable portion of regulatory resources and the lack of trust from all parties adds to the decision-making resistance. Through sufficient supervision and reasonable distribution of power and responsibility, we can establish a trusty environment, and then liberate the regulatory resources and form a virtuous circle.

4.3 Poor Coordination and the Gap in Program Operation

Limited by current hierarchical management system, there is always a 'vacuum zone' between the business coordination and convergence between government procurement. The lack of up-down communication and the guarantee of cross-sectional coordination mechanism leads to the problem in contract performance such as both supply and demand have an ambiguous understanding on the performance of the content; when the supply cannot offer the goods as demanded or both supply and demand has a different opinion on the supply standard, it shows the lack of administrative intervention function and relevant coordination mechanism. When the conflict between the two parties intensified and the problem was solved by repeated questioning and complaints, it will not only affect the efficiency of procurement, but also the credibility of government procurement. Therefore, to innovate the government procurement management mode, we should consider adding the agencies to participate in the acceptance of the contract and being responsible for coordinating the relevant functions of contract performance.

4.4 The Low Cost of Non-Compliance and the Weak Execution and Punishment

The Office of Government Procurement Management is the operational institution that accepts the government's complaints of procurement and challenges. As for more serious disciplinary problems, they can be complained to the disciplinary department and the illegal crimes should be handed over to the judicial units. However, as the Office of Government Procurement Management dealing with daily complaints, not only the professional strength of dealing with complaints is relatively weak, the right of penalties for non-compliance and breach of trust suppliers is quite limited which is deducting the tender margin and including them in the 'black list' of breach of trust. And the suppliers on the 'black-list' can simply cancel the discredited company account and re-register new companies to participate in the government procurement project bidding. This kind of low-cost and weak punishment cannot solve the problem from the root. It does not fundamentally solve the bidding violations, forging bid, accompanying - bidding and other persistent problems. From the long-term construction of government procurement, the lack of punishment mechanism, it not only represents the loss of fair in transactions of government procurement, but also does not contribute to the healthy development of government procurement.

5 Building the Model of Government Procurement Performance Evaluation Based on Stakeholders' Perspective

Based on the current situation of government procurement in China and the interests of the relevant public, it is necessary to proceed from reality, seek truth from facts, keep innovation and determine a scientific, comprehensive, reasonable and objective system of government procurement performance evaluation index system. It needs to be based on the stakeholders' perspective in order to accurately reflect the performance of government procurement, carry out a fair assessment and ensure the efficiency and effectiveness of government procurement.

5.1 Selection of the Model Indicator

To construct a government procurement performance evaluation model based on stakeholder perspective, we must first build the index system scientifically, and the selection of indicators needs to follow below principles:

a. In line with national laws, regulations and policies, constructing a complete evaluation indicator of government procurement performance which can reflect the degree and level of government procurement.
b. In line with the actual situation of government procurement in China and the procurement management of developed countries. It should be in line with the situation of our country at first which means it should be able to reflect the actual situation of the development of government procurement with the consideration of forward-looking, but also the factors that can be in line with the government procurement of other countries and regions.
c. The setting of indicators should fully reflect the situation of government procurement. The indicators should have a certain comprehensiveness and focus on the views of stakeholders eventually.
d. It should be measurable and operable, using qualitative and quantitative indicators to measure the operational status of government procurement.
e. It should be oriented and transparent which can guide the government procurement into a healthy and orderly development, which is also conducive to the establishment of a clean, efficient and transparent government.

The indicators that affect the performance of government procurement based on the above principles include economy, personnel quality, technical management guarantee, social responsibility, stakeholder evaluation and effect evaluation, as shown in Table 1.

5.2 The Selection of the Model Method

The government procurement performance evaluation should rely on the complex system of government procurement performance evaluation index based on stakeholder angle. Some indicators designed to reflect the government procurement performance are a bit abstract which is difficult to quantify [7]. Therefore, it is better to use hierarchical analysis (The analytics process, AHP method) which is according to the stakeholder-based government procurement performance evaluation indicators, design the evaluation system level model and evaluate the government procurement performance.

AHP is a semi-qualitative analysis method [8]. First of all, the decision-making problems can be seen as a large system affected by a variety of factors and these interrelated and mutually constrained factors can be ranked from high to low according to their affiliations at several levels [9]. The experts, scholars and authorities can be invited to compare the importance of every two different factors. Then they can re-use mathematical methods, sort the order of the factors layer by layer and finally analyse the ranking results to assist in decision-making.

The evaluation model of government procurement performance based on stakeholder perspective can be regarded as a function of the independent variable which includes economy, personnel quality, technical management guarantee, social responsibility, stakeholder evaluation and effect evaluation, i.e., the government procurement performance $Y = f$ (Economic, Personnel quality, Technical management guarantee, Social responsibility, Stakeholder evaluation, Effect evaluation) based on the perspective of interests.

The level model of government procurement performance assessment based on stakeholder perspective consists of three levels: (1) Target Level (G): The overall objective - the level of government procurement performance based on the stakeholder perspective. (2) Guideline level (C): includes six components of government procurement performance based on stakeholder perspective - economy (C1), quality of personnel (C2), technical management assurance (C3), social responsibility (C4), stakeholder evaluation (C5) and effectiveness evaluation (C6). (3) Assessment layer (P): includes the 17 government procurement performance assessment indicators selected earlier based on stakeholder perspective. The level model of government procurement performance evaluation based on a stakeholder perspective is shown in Table 2.

Table 1 The basic indicators that affect government procurement performance

Government procurement performance	Economy	Benefits: cost savings in government procurement are effectively realized
		Efficiency: the efficiency of government procurement has been improved.
		Transparency: Reduce the cost of supervision from stakeholders by increasing transparency in procurement
	Personnel Quality	The quality of government procurement professional: has a higher level of professionalism and morality
		The quality of Supplier: has a higher supply capacity and a higher reputation
		The quality of the stakeholders: have a high positivity and awareness of regulations
	Technical management guarantee	Technical support: a sound government procurement office equipment and information technology protection
		Management guarantee: a sound legal system to ensure the orderly conduct of government procurement activities
		Operational guarantee: reasonable operating funds to ensure the smooth development of government procurement activities
	Social responsibility	The assessment from stakeholders of the efficiency of government procurement
		The assessment from stakeholders of the effectiveness of government procurement
	Stakeholder evaluation	Stakeholders' satisfaction with government procurement
		Stakeholders' judgment on the quality of government procurement

(continued)

Table 1 (continued)

		The number of stakeholders concerned about government procurement
	Effect evaluation	Finance: Government procurement savings
		Formalization: The institutionalization and standardization of government procurement
		Innovation: Innovation of the system and mechanism of government procurement

5.3 The Calculation of the Weight of the Model Indicator

Through a large number of questionnaires and expert interviews, a comprehensive assessment resulted in a government procurement performance evaluation matrix based on a stakeholder perspective, as shown in Table 3.

By using the hierarchical analysis software Yaahp processing the data obtained, it shows the result of the relative importance of each stakeholder-based government procurement performance assessment indicator relative to its previous level indicator, i.e. the weight, as shown in Table 4.

CI = 0.0550; RI = 1.24; CR = 0.0443 < 0.1, the judgment matrix has good consistency (Table 5).

CI-0; RI-0.58; CR-0-lt; 0.1, the judgment matrix has good consistency (Table 6).

CI-0; RI-0.58; CR-0-lt; 0.1, the judgment matrix has good consistency (Table 7).

CI-0.0268; RI-0.58; CR-0.0462-lt; 0.1, the judgment matrix has good consistency (Table 8).

CI-0; RI1E-6; CR-0 < 0.1, the judgment matrix has good consistency (Table 9).

CI-0.0268; RI-0.58; CR-0.0462-lt; 0.1, the judgment matrix has good consistency (Table 10).

CI-0; RI-0.58; CR-0-lt; 0.1, the judgment matrix has good consistency.

$$\begin{aligned} CI\sum_{i=1}^{6} C_i CI_i &= 0.1587 \times 0 + 0.214. \times 0 + 0.1414 \\ &\quad \times 0.0268 + 0.1782 \times 0 + 0.0833 \times 0.0268 \\ &\quad + 0.2245 \times 0 \\ &= 0.0060 \end{aligned}$$

$$RI = \sum_{i=1}^{6} C_i RI_i = 0.1587 \times 1.24 + 0.2140 \times 0.58$$

Table 2 A hierarchical model of government procurement performance evaluation based on a stakeholder perspective

Government procurement performance (G) based on stakeholder perspective	Economic C1	Benefit P1: cost savings in government procurement are effectively realized
		Efficiency P2: the efficiency of government procurement has been improve
		Transparency P3: Reduce the cost of supervision from stakeholders by increasing transparency in procurement
	Personnel quality C2	The quality of government procurement professional: has a higher level of professionalism and morality
		Supplier quality P5: has a higher supply capacity and a higher reputation
		The quality of relevant stakeholders P6: have a high positivity and awareness of regulations
	Technical management guarantee C3	Technical support P7: a sound government procurement office equipment and information technology protection
		Management guarantee P8: a sound legal system to ensure the orderly conduct of government procurement activities
		Operation Guarantee P9: There are reasonable operating funds to ensure the smooth development of government procurement activities
	Social Responsibility C4	P10 Stakeholder Assessment of the Efficiency of Government Procurement
		P11 Stakeholder Assessment of the Effectiveness of Government Procurement
	Stakeholder evaluation C5	P12 Stakeholder satisfaction with government procurement
		P13 Stakeholder's judgment on the quality of government procurement

(continued)

Table 2 (continued)

		Number of P14 stakeholders concerned about government procurement
	Effect evaluation C6	P15 Finance: Procurement costs saved by government procurement
		P16 Formalization: Institutionalization and Standardization of Government Procurement
		P17 Innovation: Innovation of The System and Mechanism of Government Procurement

Table 3 The meaning of the 9th order scale

Serial number.	a_{ij}	Importance, etc.
1	1	The element *i* is as important as the element *j*
2	3	The element *i* is slightly more important than the element *j*
3	5	The element *i* and the element *j* are significantly important
4	7	The element *i* is strongly important with the element *j*
5	9	The element *i* and element *j* is extremely important
6	1/3	The element *i* is slightly less important than the element *j*
7	1/5	The element *i* and the element *j* are obviously not important
8	1/7	The element *i* is strong lying with the element *j*
9	1/9	The element *i* and element *j* are extremely unimportant

$$\begin{aligned} &+ 0.1414 \times 0.58 + 0.1782 \\ &\times 0 + 0.0833 \times 0.58 \\ &+ 0.2245 \times 0.58 \\ &= 0.5814 \end{aligned}$$

$$CR = \frac{CI}{RI} = \frac{0.0060}{0.5814} = 0.0104 < 0.1$$

The total judgment matrix has good consistency and the weight values obtained by data processing are valid.

Table 4 The relative importance of criterion-level capability index C to target level G

Government Procurement Performance G	Economy C1	Personnel Quality C2	Technical management guarantee C3	Social Responsibility C4	Stakeholder evaluation C5	Effect evaluation C6
Economy C1	1	1/2	1	1	2	1
Personnel Quality C2	2	1	1	2	3	1/2
Technical management guarantee C3	1	1	1	1/2	2	1/2
Social Responsibility C4	1	1/2	2	1	2	1
Stakeholder evaluation C5	1/2	1/3	1/2	1/2	1	1/2
Effect evaluation C6	1	2	2	1	2	1
Single layer weight W	0.1587	0.2140	0.1414	0.1782	0.0833	0.2245

Table 5 The relative importance of evaluation level to criterion level C1

Economy C1	P1	P2	P3
P1	1	1	1/2
P2	1	1	1/2
P3	2	2	1
Single layer weight *W*	0.2500	0.2500	0.5000

Table 6 The relative importance of evaluation level to criterion level C2

Economy C2	P4	P5	P6
P4	1	2	2
P5	1/2	1	1
P6	1/2	1	1
Single layer weight *W*	0.5000	0.2500	0.2500

Table 7 The relative importance of evaluation level to criterion level C3

Technical management guarantee C3	P7	P8	P9
P7	1	4	2
P8	1/4	1	1
P9	1/2	1	1
Single layer weight *W*	0.5842	0.1840	0.2318

Table 8 The relative importance of evaluation level to criterion level C4

Social responsibility C4	P10	P11
P10	1	4
P11	1/4	1
Single layer weight *W*	0.8000	0.2000

Table 9 The relative importance of evaluation level to criterion level C5

Technical management guarantee C5	P12	P13	P14
P12	1	3	3
P13	1/3	1	2
P14	1/3	1/2	1
Single layer weight *W*	0.5936	0.2493	0.1571

Table 10 The relative importance of evaluation level to criterion level C6

Technical management guarantee C6	P15	P16	P17
P15	1	1	2
P16	1	1	2
P17	1/2	1/2	1
Single layer weight *W*	0.4000	0.4000	0.2000

5.4 Empirical Research

Based on the network public information, questionnaire and interview data, the C City 2018 government procurement performance is selected as the research object in this essay and the relevant performance data has been collected. The score range is 1, 3, 5, 7, 9 which respectively representing very good, good, general, poor, very poor. The data table processed based on survey data is shown in Table 11.

Overall, the overall score of City C procurement performance in 2018 was 6. 81 which is between 7 and 5. The score is medium-high indicating that the general situation is quite good. In order to improve the performance of government procurement, it is necessary to further improve some parts. For example, the score of economic, personnel quality, social responsibility and effectiveness evaluation are lower than 7 and these work needs to be further improved to improve the overall performance of government procurement.

Table 11 City C 2018 Government Procurement Performance Assessment

Target level	Score	Criterion level	Score	Evaluation level	Score
Government procurement performance (G) based on stakeholder perspective	6.81	Economy C1	6.38	P1 P2 P3	6.92 7.31 5.39
		Personnel Quality C2	6.61	P4 P5 P6	6.93 7.35 4.72
		Technical management guarantee C3	7.19	P7 P8 P9	7.05 7.95 6.32
		Social Responsibility C4	6.61	P10 P11	6.59 6.06
		Stakeholder evaluation C5	7.45	P12 P13 P14	7.82 6.37 6.84
		Effect evaluation C6	6.99	P15 P16 P17	6.62 7.17 6.75

References

1. J. Wang, Exploration on the management of state-owned assets in universities. Exp. Technol. Manage. **33**(06), 248–251 (2016)
2. Fan Peng, Problems existed in management of state-owned assets in institutions of higher education and countermeasures. J. Macroeconomic Manage. **S1**, 74–75 (2017)
3. Y. Xu, Z. Fan, T. Chen, Exploring the supply-side reform of state-owned assets management in universities. Laboratory Res. Exploration **35**(11), 254–256 + 291 (2016)
4. E.I. Papageorgiou, P.P. Groumpos, A new hybrid method using evolutionary algorithms to train Fuzzy Cognitive Maps. Appl. Soft Comput. J. (4) (2014)
5. Cho, Y. Park, *Expert Systems With Applications* (1) (2017)
6. W.-W. Wu, Choosing knowledge management strategies by using a combined ANP and DEMATEL approach. Expert Syst. Appl. (3) (2017)
7. R. Yu, G.-H. Tzeng, A soft computing method for multi-criteria decision making with dependence and feedback. Appl. Math. Comput. (1) (2016)
8. Valuation of urban industrial land: an analytic network process approach. Eur. J. Oper. Res. (1) (2016)
9. J.-R. Yu, S.-J. Cheng, An integrated approach for deriving priorities in analytic network process. Eur. J. Oper. Res. (3) (2016)

Non-financial Influencing Factors of Software Enterprise Credit Level

Lingling Wang, Shifeng Liu, and Shouting Miao

Abstract Enterprises are the cornerstone of the market economy, and the lack of corporate credit has seriously affected country's social and economic development. With the advent of the era of big data, corporate credit information is constantly enriched and refined, making it possible to eliminate corporate credit rating defects and improve personalized industry credit evaluation. This paper uses 500 software companies as samples, through factor analysis and regression analysis, the non-financial impact factors of the software industry's credit level are empirically evaluated, which provides decision-making reference for government supervision and research analysis, and provides useful support for the establishment of a software industry-specific enterprise credit evaluation system.

Keywords Corporate credit level · Non-financial factors · Software enterprises

1 Introduction

With the development of the market economy, the construction of a social credit system has become one of the symbols of China's socialist market economy moving towards maturity, and is an important means of improving the socialist market economic system and stabilizing social and economic development. At present, the problem of lack of credit has become a "bottleneck" restricting China's economic development. According to relevant statistics, China's annual losses caused by the lack of corporate credit amount to 600 billion yuan, and the bad debt rate is as high as 1%–2%. By comparison, the typical bad debt rate for enterprises in mature market

L. Wang (✉) · S. Liu
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 18120622@bjtu.edu.cn

S. Liu
e-mail: shfliu@bjtu.edu.cn

S. Miao
China State Railway Group Co., Ltd, Beijing, China
e-mail: miaoshouting@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_14

economies is 0.25%–0.5%. In addition, of the approximately 4 billion contracts signed each year, the compliance rate is only 50%. The legal bottom line is no longer able to curb business violations such as corporate breach of contract and fraud. The construction of the corporate credit system has become particularly important in the process of China's economic development.

As a software company that has gradually become the backbone of economic and social transformation, it is gradually becoming a leading force in the transformation of China's industrial economy. Therefore, the operating conditions and credit performance of software companies have gradually become an important part of the construction of the national enterprise credit system. In order to enhance the awareness of the integrity of software companies and create a good credit environment for the software industry, the study of factors affecting the credit level of software companies has gradually attracted the attention of scholars.

At present, corporate credit evaluation mostly determines the credit level of an enterprise based on its financial status, that is, its credit ability, and the research on the financial influence factors of corporate credit has been in-depth and mature [1]. Similarly, in the software industry, most credit rating agencies also focus on the financial indicators of enterprises. However, financial indicators can only represent the past operating results of an enterprise, and cannot fully reflect the expected solvency of the enterprise [2]. Therefore, studying the impact of non-financial factors such as corporate organizational structure and social performance on corporate credit can objectively evaluate the corporate credit level from multiple perspectives, and with the integration of corporate credit data, companies have sufficient capabilities to deeply analyze the enterprise Non-financial factors of credit.

2 Theoretical Basis and Hypothesis

Since the 1990s, research on the non-financial influence factors of corporate credit has gradually entered everyone's horizons, and many scholars have done a lot of research on methods and models of corporate credit evaluation. Indicators are the basis for conducting research on forecasting and evaluation. Therefore, scholars pay great attention to the study of enterprise evaluation index system [3]. Brown proposed that the five aspects of corporate culture, interpersonal relationship, organization, society and nature will affect the credit level of the enterprise [4]. Shu Yan proposed to include corporate credit reporting and personal credit reporting of executives into the system indicators [5]. Duan Wei calculated the objective weights of the indicators based on the entropy weight method, used the G1 method to obtain the subjective weights of the indicators, and borrowed the game theory method to propose that the enterprise's corporate individual factors and macro-environmental factors have significant effects on corporate credit evaluation [6]. Wang Wenrong and others believe that in the context of big data, the evaluation of the credit situation of small and medium-sized enterprises should integrate the three aspects of corporate credit, corporate credit, and corporate social credit [7].

It is concluded from the above that there is no unified and authoritative conclusion for non-financial indicators in corporate credit evaluation systems. Based on the research results of scholars on the factors affecting corporate credit, this paper intends to explore the non-financial impact factors of corporate credit from four aspects: management scale, innovation ability, financing ability and operating risk.

2.1 Management Scale Factor

The management of an enterprise refers to the personnel with management responsibility in the enterprise. It is an element that gives life and vitality to the enterprise and is responsible for various decisions and action directions of the enterprise. As stated by Poster, enterprise managers, especially senior managers, are the main decision makers of the enterprise. Compared with other members of the organization, the company's senior managers have greater influence on the company's decision-making and have more opportunities to influence the company's business strategy. And operating conditions [8]. The difference between enterprises in fulfilling social responsibilities and obligations is largely determined by the attitudes of corporate managers [9], and corporate social responsibility can affect the debt default risk and credit rating of enterprises [10]. Personal attitudes can affect a company's default risk and credit rating to a certain extent. In addition, Zheng Haiyuan and other studies have confirmed that management risk appetite has a significant correlation with corporate credit ratings [11], but when a manager is in a very strong position, he will insist on his ideas, whether correct or not [12]. It can be considered that when the scale of the management of an enterprise is unreasonable, authoritarian situations are prone to occur, and certain tendencies are likely to occur when making decisions, affecting the normal operation of the enterprise. Therefore, a reasonable management scale can reduce the possibility of corporate dishonesty and promote the company to maintain a good credit level. Based on the above analysis, this article proposes hypotheses:

H1: The corporate credit level is significantly positively related to the size of the management.

2.2 Creative Ability Factor

Innovation is the source of power for economic and social development. As the backbone of the development of information technology, software companies need to strengthen their innovative position and leading role. Intellectual property rights, as a result of enterprises' pursuit of innovation, are the exclusive rights for enterprises to protect their intellectual labor, and are also good performances for enterprises to shoulder the important task of social development. Regarding the impact of innovation ability on the credit level of enterprises, some scholars have proposed that for innovative enterprises, a credit evaluation system with intellectual property as the core should be established [13]. Good intellectual property credit is conducive to cooperation, expansion and market promotion of enterprises. Some scholars have pointed out that the innovation ability composed of two indicators of R&D expenditure input and R&D personnel ratio is related to the credit level of the enterprise [14], but the R&D expenditure investment and the ratio of R&D personnel do not represent the results obtained by the enterprise. It's not that the more the cost is invested, the more gains will be obtained. Therefore, this article takes software companies as the research object, most of the enterprises are classified as innovative enterprises, this article takes the amount of intellectual property as a measure of the enterprise's innovation ability, measures the enterprise's innovation ability by the achievements obtained by the enterprise, and according to the definition of intellectual property Three indicators were refined, including the number of software copyrights, the number of copyrights and the number of trademarks. Based on the above analysis, this article makes assumptions:

H2: Corporate credit level is significantly positively related to innovation ability.

2.3 Financing Capacity Factors

Financing ability is the ability of an enterprise to raise funds, and it is the embodiment of an enterprise's ability to produce, operate and grow. Regarding the impact of corporate financing on corporate credit, some scholars have proposed that, from the perspective of supply chain finance, the construction of corporate credit risk assessment systems mainly covers macro-environmental conditions, financing conditions, core reputation, and cooperation between supply chain synergy partners. The situation waits for four levels [15]. The better the financing status, the stronger the financing ability, the more stable the financial status of the enterprise, the lower the credit risk, and the higher the credit level.

At present, pledge of equity as a fund-raising behavior has been widespread in China's capital market, especially in service industry companies such as software companies. Because there are not enough fixed assets as collateral, shareholders within the enterprise will take the form of equity pledge. Financing. However, some studies have pointed out that when the number of shareholders is small and the proportion of equity is concentrated, due to the relatively strong control rights, the controlling shareholder will continue to strengthen the hollowing out of listed companies and the appropriation of other shareholders' interests for the purpose of embezzling funds. Meeting their own financing needs has led to increasing levels of corporate financing constraints [15]. Increased financing constraints will directly affect the financial status and credit risk of the enterprise. Therefore, the number of shareholders of an enterprise affects the concentration of equity ratios. When the equity ratios are relatively concentrated, it is easy to increase the level of capital constraints, which in turn affects the financial status of the enterprise, and ultimately leads to an increase in corporate credit risk and a reduction in credit level. Based on the above analysis, this article proposes hypotheses:

H3: The corporate credit level is significantly positively related to the number of shareholders.

2.4 Operational Risk Factors

Operational risk refers to the possibility that an enterprise's future revenue will be damaged due to uncertain factors in the production and operation process and the value of the enterprise will change. The occurrence and accumulation of risk factors will increase the business risk of the enterprise, increase the possibility that the revenue of the enterprise will be damaged, and then threaten the normal production and operation activities of the enterprise. In recent years, research on corporate credit influencing factors has led many scholars to propose that operational risk factors have important impacts on corporate credit levels in different fields. Gejirifu De built a credit risk evaluation index system for power retail enterprises from the perspective of sustainable development of power retail enterprises. One of the important indicators is the operating credit risk [16]. Meng Bin and Yang Yue et al., take small construction enterprises as their research subjects and believe that corporate legal disputes as business risks will affect their credit level [17]. There are also scholars who incorporate the characteristics of simulation internships and include corporate non-compliance, dishonesty and breach of contract as business risks into the credit evaluation model [18]. It can be found from the existing research that the determination of operational risk factors is subjective, and the rationality of these standards needs to be explored. Therefore, this article adopts the number of public judicial announcements as an indicator of enterprise operating risk.

In addition, for software companies, the outbreak of negative online public opinion information should also belong to the category of operating risk factors, because online public opinion is a mapping of social hot events in the Internet space. When a more prominent bad event occurs in an enterprise, online public opinion spreads The breadth and speed of the company can quickly reflect related issues, and provide an early warning to the company's operating risks. Studies have pointed out that efficient Internet public opinion monitoring can provide early warning of risks to the public opinion of SMEs and the government [19]. The impact of negative online public opinion on corporate credit has been analyzed by scholars on the "Longevity Biological Vaccine Incident", and pointed out that the generation of negative public opinion not only caused serious damage to the direct profitability of enterprises, but also caused the company's level of financing cash flow to drop sharply. Horizontal "cliff-like" declines [20]. Therefore, negative public opinion as a result of business management issues is an important index factor to determine business risk. According to the dissemination characteristics of public opinion on the Internet, using the negative public opinion of the company as an influencing factor can not only reflect the credit situation of the company in the past, but also play a certain early warning effect on the credit situation of the company [21–23]. In summary, this article proposes hypotheses:

H4: Significant negative correlation between corporate credit level and corporate operating risk.

3 Empirical Design

3.1 Test Line

This paper intends to apply the statistical methods of factor analysis and multiple linear regression analysis to establish the corresponding regression models to study the relationship between management scale, innovation ability, financing ability, business risk and corporate credit level to verify whether the above assumptions hold. The testing idea of this paper is mainly divided into three steps: First, collect the corresponding sample data according to the set index variables, and perform the corresponding descriptive analysis on the collected sample data; then, use the factor analysis method to analyze the collected data. The sample variables are subjected to dimensionality reduction processing to test the rationality of the hypotheses initially. Finally, a multiple regression model of factors, explanatory variable samples, and corporate credit is established to obtain the correlation between the explanatory variables and the explained variables, and then scientifically perform hypothesis testing. And analyze the degree of influence of explanatory variables on the explanatory variables.

3.2 Research Sample

This paper takes software companies as the research object, selects 500 representative software companies, and uses related tools and manual retrieval methods to obtain sample data from open credit reference websites. The sample data contains a total of 10 attribute dimensions, including 9 independent variables and 1 dependent variable. The setting of the variables is shown in Table 1.

3.3 Descriptive Statistics of Sample Data

Due to the different sources of sample data in this article, the nature of each indicator is different, and there is a certain dimension between the data. In order to remove the unit limitation of the data, it is converted into dimensionless data. In particular, when the evaluation index processing is involved, data standardization is necessary. One of the selected data processing methods. Among them, the most typical data standardization processing method is normalization processing to improve the model's convergence

Table 1 Table of variables

Variable type	Factor classification	Variable	Indicator
Dependent Variable			Corporate Credit Score
Independent variables	Management size	X_1	Number of key management personnel
	Innovation Ability	X_2	Number of Trademarks
		X_3	software copyrights
		X_4	number of copyrights
	Financing capacity	X_5	Number of shareholders
	Operational risk	X_6	Number of referee documents
		X_7	Number of court announcements
		X_8	Number of hearing announcements
		X_9	Negative public opinion

speed and model accuracy. The normalization process can normalize all the original data to the [0, 1] interval. Since the specific values of each sample data are known, this paper uses the min–max normalization method for normalization processing. The conversion function is as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Among them, $x_{\min}$ is the minimum value of any dimension variable data, and $x_{\max}$ is the maximum value of any dimension variable data.

In this paper, descriptive statistics are made on the explanatory variables and the data of the explained variables of 500 software enterprises in order to achieve a preliminary analysis of the concentration trend and the degree of dispersion of the sample values of each variable. The results from the statistical analysis are shown in Table 2.

Table 2 Table of descriptive statistical results

Indicator	Variable	Mean	Standard deviation
Number of key management personnel	X_1	0.2078	0.1440
Number of Trademarks	X_2	0.0204	0.0680
Software copyrights	X_3	0.1061	0.1324
Number of copyrights	X_4	0.0067	0.0556
Number of shareholders	X_5	0.0608	0.1222
Number of referee documents	X_6	0.0139	0.0603
Number of court announcements	X_7	0.0156	0.0614
Number of hearing announcements	$X8$	0.0088	0.0475
Negative public opinion	X_9	0.0123	0.0598
Corporate Credit Score	Y	0.6782	0.2220

As can be seen from the standard deviation of the sample data of the variables in the table above, the standard deviation of the corporate credit score is 0.2220, and the degree of dispersion is relatively large, indicating that the credit level of the selected sample companies is significantly different, which is more conducive to the analysis of explanatory variables. Correlation and degree of influence; the standard deviation of the number of main management personnel of the company is 0.14, and the standard deviation of all explanatory variables is the largest, indicating that there is a large difference in the size of the management layer among the sample companies, reflecting that the sample has a strong The standard deviation of the announcement of the hearing is 0.047. Among the explanatory variables, the standard deviation is smaller, indicating that the data of this variable has less fluctuation among the sample companies.

4 Empirical Test and Analysis

4.1 *Factor Analysis*

Aiming at the two influencing factors of enterprise innovation ability and enterprise operation risk, this paper has obtained a total of 7 index variables, and each type of impact factor involves multiple index variables, and the index variables involved are relatively large, which makes the model have certain multicollinearity risks. It affects the subsequent analysis, so this article uses SPSS software to use factor analysis to reduce the dimension of the explanatory variables.

Seven explanatory variables of the number of software copyrights, copyrights, trademarks, number of court announcements, number of judgment documents, number of court announcements, and number of negative public opinions were entered into the variables of factor analysis. The value of KMO statistics was 0.627 > 0.5. Factor analysis is still available. According to the factor analysis results output by the SPSS software, it can be known that the input explanatory variables are finally extracted with 2 factors, and the cumulative variance contribution rate of the 2 factors is 72.367%, and the information interpretation degree of the original variables is higher than 60%, indicating that the factors affect the variables. The ability to explain is acceptable. Two factors were extracted according to the criterion of "initial eigenvalue" greater than 1. The difference between the variance contribution rate of the two factors after rotation was reduced, and the cumulative variance contribution rate was the same. Relatively speaking, the factor interpretation ability was relatively good. The interpretation results of the total variance of the factor analysis are as follows (Table 3).

The component matrix of the explanatory variable is obtained by factor analysis, and the constituent elements of the two factors can be seen from the matrix results. Factor 1 can reflect the company's innovation ability, and the variable indexes that meet the factor load limit include three explanatory variables: software copyright

Table 3 Table of descriptive statistical results

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.172	51.475	51.475	1.548	50.955	50.955
2	1.042	20.892	72.367	1.079	21.412	72.367

Table 4 Table of composition matrix aft

	Component	
	1	2
Softeware copyrights	0.780	
Number of trademarks	0.794	
Number of copyrights	0.668	
Number of hearing announcements		0.905
Negative public opinion		0.899
Number of referee documents		0.786
Number of court announcements		0.496

number, copyright number, and trademark number; factor 2 can reflect the business risk of the enterprise and satisfy the variable load limit The indicators include four variables: the number of announcements of court hearings, the number of adjudication documents, the number of court announcements, and the number of negative public opinions. The rotated component matrix is shown in the Table 4.

4.2 *Multiple Regression Analysis*

In order to test the hypotheses proposed in this article, and to judge the relationship between each indicator variable and the credit level of the enterprise, two factors obtained from the factor analysis, the number of shareholders X_1, the number of key management personnel X_2, and the explanatory variables are used as the explanatory variables. Perform multiple regression analysis. The regression analysis between variables was performed by SPSS software, and the multicollinearity problem of the model was tested. The statistical results are as follows (Table 5).

From the analysis of variance table, it can be seen that the significance of the F test (p-value) $= 0.000 < 0.05$, that is, the model can be considered to be at the significance level of 0.01. The linear relationship established by the explanatory variable's corporate credit level has significant statistical significance. It can be said that the number of shareholders, factor 1, factor 2 and the number of key management personnel have a significant linear effect on the corporate credit level (Table 6).

Table 5 Table of analysis of variance

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig
1	Regression	5.244	4	1.311	33.533	0.000[b]
	Residual	19.354	495	0.039		
	Total	24.599	499			

[a] Dependent Variable: Corporate Credit Score
[b] Predictors: (Constant), Factor 1, Factor 2, Number of Trademarks, Number of key management personnel

The multiple regression equation can be obtained from the multiple regression coefficient table:

$$Y = 0.057 + 0.454\,X_1 + 0.191\,X_5 + 0.045\,\mathrm{Factor1} - 0.021\,\mathrm{Factor2}.$$

After t-test, the significance (p-value) of the partial regression coefficients of each explanatory variable was less than 0.05, indicating that the coefficients of each explanatory variable were not 0, and there was a significant correlation between each explanatory variable and the explained variable. In addition, a VIF test was performed in this article. The results show that the VIF values of each explanatory variable are less than 5, indicating that there is no problem of multicollinearity between regression models.

It can be seen from the partial regression coefficients in the table that the number of key management personnel X_1, the number of shareholders X_5, and the company's innovation ability (Factor1) are positively related to the company's credit level. Operational risk (Factor2) has a negative correlation with the corporate credit level. The larger the number of operating risk indicators, the lower the corporate credit level. Therefore, the results are consistent with the assumptions in this paper. From the standard partial regression coefficient, it can be seen that the standard partial regression coefficient of the number of key management personnel X_1 is the largest, that is, its degree of influence on the corporate credit level is relatively greater; the degree of influence of the remaining explanatory variables is innovation ability, number of shareholders and enterprise Operating risk, relatively speaking, corporate operating risk has the least impact on corporate credit.

5 Conclusion

This paper takes software companies as the main research body. The research finds that some non-financial factors of companies can have a significant impact on their credit level. Through empirical research and analysis, the following conclusions are

Table 6 Table of regression coefficient

Coeffients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	0.057	0.016		35.544	0.000		
	Number of shareholders	0.191	0.084	0.105	2.278	0.023	0.749	1.335
	Number of key management personnel	0.454	0.073	0.294	6.244	0.000	0.717	1.395
	Factor1	0.045	0.009	0.201	4.836	0.000	0.923	1.084
	Factor2	−0.021	0.009	−0.094	2.358	0.019	0.995	1.005

[a] Dependent Variable: Corporate Credit Score

drawn: First, factor analysis was used to extract innovations from the seven explanatory variables of the number of software copyrights, the number of copyrights, the number of trademarks, the number of court announcements, the number of court documents, the number of court announcements, and the number of negative public opinions. Two factors: ability and operating risk; secondly, the size of the management, the ability to innovate, the financing ability, and the operating risk factor can be significantly affected by the multiple linear regression method to test the creditworthiness of the company. Number of managers, innovation ability factor, number of shareholders, and operating risk factor.

The research results show that the non-financial factors in the software industry can reflect the credit level of enterprises to a certain extent. In terms of enterprise management scale, a good corporate organizational structure can more fully formulate corporate strategies and decisions, and assist in the efficient implementation and implementation of decisions; a healthy corporate organizational structure is beneficial to balance the power of various decision makers and coordinate relevant stakeholders The relationship between them, standardizing the daily operation behavior of the enterprise, forming good corporate operating habits and corporate culture, making the internal management of the enterprise relatively stable, thereby improving the credit level of the enterprise. In terms of enterprise innovation capability, innovation capability is the core of the economic competition of the enterprise, so that the enterprise has sufficient development potential. The strong innovation ability can continuously provide economic value to the enterprise, stabilize the operation of the enterprise, and improve the credit level of the enterprise. In terms of financing capacity, appropriately increasing the number of shareholders can diversify equity to a certain extent, reduce the funding risk brought by financing constraints, help alleviate corporate financing barriers, reduce corporate credit risk due to decline in performance and tight capital, and improve corporate credit Level. In terms of business risks, judicial information and negative public opinion information can to some extent reflect the business problems of the enterprise over a period of time, especially negative public opinion information. The negative public opinion can sensitively and quickly capture some potential and emerging risks of the enterprise. Furthermore, it reflects the credit risk of the enterprise, which negatively affects the credit level of the enterprise. However, from the perspective of the impact, the impact of operational risk on corporate credit is relatively small, indicating that risk factors have not been given sufficient attention, and a reasonable supervision mechanism needs to be established to promote enterprises to avoid related risks.

In view of the limited data acquisition, this article has certain limitations: first, this article only conducted an inquiry and analysis on some non-financial factors, and the stability of the research results needs further exploration and testing; second, this article only includes relevant quantitative data into the analysis, without considering Related qualitative variables such as personnel quality; Finally, this article only analyzes the non-financial factors of software company credit, and the universality of the research results needs further research.

References

1. X. Luo, Y. Song, H. Xu, Analysis on non-financial influencing factors of enterprise credit. Sci. Technol. Manage. **17**(04), 59–66 (2015)
2. D. Liandi, Z. Rong, Thinking on the construction of credit evaluation index system for SMEs. Qi Luzhutan **06**, 9–11 (2017)
3. H. Song, Li. Xiaojun, A summary of research on chinese enterprise credit in the context of big data: an analysis based on CSSCI retrieval papers. Financial Theor. Pract. **10**, 107–113 (2018)
4. M. Brown, Corporate integrity and public interest: a relational approach to business ethics and leadership. J. Bus. Ethics **66**(3), 11–18 (2006)
5. S. Yan, Research on the construction of credit rating index system for small and micro enterprises. Financial Theor. Pract. **05**, 105–108 (2015)
6. D. Ye, Research on credit evaluation of small and micro enterprises based on game theory combination empowerment. Credit Information **37**(09), 12–17 (2019)
7. W. Wang, Y. Chen, Construction of credit evaluation system for small and medium-sized enterprises under the background of big data. .E-commerce (07), 46–47 (2018)
8. J.E. Post, T.L. Anne, W. James, *Business and Society: Corporate Strategy and Public Policy and Ethics*. Zhang Zhiqiang, Trans. (Renmin University of China Press, Beijing, 2005)
9. T. Xue, Research on the social responsibility attitude of senior managers in private enterprises and its influencing factors. East China Econ. Manage. **29**(03), 41–45 (2015)
10. L. Hui, Z. Shichen, Corporate social responsibility and corporate credit risk assessment: innovation of logistic model based on stakeholder perspective. J. Hunan Univ. (Soc. Sci. Edn.) **33**(02), 88–95 (2019)
11. H. Zheng, Y. Du, Management risk appetite, earnings management and corporate credit rating. Finance Accounting Monthly (12), 60–69 (2019)
12. U. Malmendier, G. Tate, J. Yan, Overconfidence and early-life experiences: the effect of managerial traits on corporate financial policies. J. Finance **5**, 1687–1733 (2011)
13. Li. Xiaoyang, Analysis on the establishment of knowledge value credit evaluation system. Exploration **05**, 185–190 (2017)
14. H. Qin, Construction of credit risk evaluation index system for SMEs from the perspective of supply chain finance. Time Finance **09**, 171–172 (2019)
15. X. Sun, Z. Yu, Research on financing constraints and financing models of China's service industry enterprises: an empirical analysis based on panel data of listed companies in the service industry [J/OL]. J. Dalian Univ. Technol. (Soc. Sci. Edn.) (02) [2020–03–03] (2020). https://doi.org/10.19525/j.issn1008-407x.2020.02.004
16. G. De,Z. Tan, M. Li, et al., A credit risk evaluation based on intuitionistic fuzzy set theory for the sustainable development of electricity retailing companies in China. Energy Sci. Eng. (1) (2019)
17. M. Bin, Y. Yue, D. Yijie, Research on credit evaluation model of small construction enterprises based on significantly differentiating two types of customers. Syst. Eng. Theor. Pract. **39**(2), 346–359 (2019)
18. X. Su, P. Wei, *Research on Credit Ratings of Commercial Banks to Enterprises in School Simulation Practices*
19. K. Liu, K. Ma, Z. Yue, Analysis and design of public opinion pre-warning analysis platform based on vertical search engine, in *2017 IEEE 14th International Conference on e-Business Engineering (ICEBE)*, Shanghai (2017), pp. 288–292
20. N. Erbao, A study on the impact of negative internet public opinion events on the credit level of general industrial and commercial enterprises. Credit Information **37**(01), 16–20 (2019)

21. Z. Ma, S. Xiao, Closed form valuation of vulnerable European options with stochastic credit spreads. Econ. Comput. Econ. Cybern. Stud. Res. **53**(4), 293–311 (2019)
22. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik-Technical Gazette **25**(2), 528–535 (2018)
23. L.M. Wang, Z.Y. Hao, X.M. Han, R.H. Zhou, Gravity theory-based affinity propagation clustering algorithm and its applications. Tehnicki Vjesnik-Technical Gazette **25**(4), 1125–1135 (2018)

Evacuation Design and Simulation of Rail Transit Platforms

Fang Wang

Abstract As one of the most important public transportation facilities in the city, urban rail transit facilities meet a large number of travel needs of the public, however, they are not the most important public transportation facilities in the city. Since most of the rail transit facilities are underground spaces, they are characterized by large flows of people and confined spaces, which can easily cause major disasters in the event of a disaster. Personnel injuries, therefore, are especially important for the study of passenger disaster evacuation in rail transit. The thesis proposes a theory of passenger evacuation in urban rail transit hubs based on system simulation, and studies the platform structure, platform structure, and platform structure of rail transit hubs. The influences of crowd size and crowd density on passenger evacuation, and the design of optimal evacuation scenarios for emergencies through simulation experiments. Evacuation paths and methods. Simulation results show that the evacuation strategy combining the three factors produces good results and can solve the bottleneck problem of emergency evacuation at the station. Reducing casualties, providing guidance for the design of emergency evacuation plans for urban rail stations, and providing guidance for disaster evacuation strategies in stations. Provide informative advice on design.

Keywords Rail transit · Passengers · System simulation · Evacuation

1 Introduction

The rapid development of urban rail transit has alleviated the current situation of urban transportation to a certain extent and provided urban residents with a fast and convenient transportation mode, but also exposed some problems in terms of operational safety. As the fundamental unit of the urban rail transit system, a large number of passengers gather, transfer and evacuate here, and the flow of people is very large. In addition, the overall structure of urban rail transit stations is usually

F. Wang (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: wf_smile_ff@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_15

located underground, and the space is closed and relatively small. In the event of fire, terrorist attacks, large passenger flows and other emergencies, emergency evacuation of passengers is extremely difficult, and it is easy to cause major casualties. Because urban rail transit stations are located underground and the space is relatively small and closed, the layout of the station and the organization of passenger flow will directly affect the difficulty of evacuation of a station fire accident or explosion. If it is not evacuated in time, it is likely to cause secondary accident such as stomping inside the station, resulting in a large number of casualties and seriously endangering the lives and property of passengers. How to effectively crowd people in the case of evacuation and prevent mass deaths and injuries has emerged as an important research topic in the field of public safety at home and abroad, as well as an important component of urban rail design.

2 Literature Review

The simulation of urban rail transit evacuation is one of the current good research directions, and many scholars have studied the simulation model. Simulation is one of the most effective ways to study the evacuation of urban rail transit.

Ma et al. (2009) Considering that the evacuation of people leaving the building is affected by many factors such as the emergency environment, the composition of the building and the behavior of the evacuated people, it is very important to determine the impact mechanism and degree to ensure safe evacuation, the crowd evacuation at the subway station was simulated, and the impact parameters were quantitatively analyzed, including walking speed, etc. [1]. Dai et al. (2009) introduced a visual simulation system for the evacuation of people at subway stations. Based on the analysis of fire simulation values and evacuation simulation values, a data fitting algorithm was designed and the real-time visual output problem was solved by it. The system was reviewed from the aspects of model configuration, function, model environment establishment, and visual simulation development tools [2]. Kang et al. (2011) proposed a stochastic model for emergency evacuation simulation of subway stations. The model includes many random variables such as occupant load, initial occupant load and pre-evacuation time. Random events of fire include fire location, fire growth rate. The influence of smoke can be used to determine the accessibility of the exit path [3]. Han et al. (2012) simulated the crowd evacuation of a subway train fire, set density of four passengers in the train, and compared the software simulation results based on the calculation results of the specifications, reflecting the relatively reliable impact of passenger density on crowd evacuation [4]. Yang et al. (2013) studied the Fire-Human Body Model FDS + Evac. This model is a decentralized design method (DDM) for fire emergency evacuation simulation that reduces simulation time and cost. Currently this method is widely used in subways, mainly Study the impact of multiple factors on fire emergency evacuation [5]. Zhao et al. (2014) proposed an evacuation agent-centric simulation model in order to analyze evacuation behavior and optimize evacuation strategies for rail transit systems. First, by considering the attributes, status and decision-making behaviors of evacuees, an evacuation agent model is established, and the operation principle and construction process of the multi-agent simulation model are discussed [6]. Cai et al. (2016) comprehensively studied the fire spread and personnel evacuation in a deep buried subway model. The model of STEPS software was used to obtain the evacuation rules. The FDS + Evac program was used to simulate evacuation in a fire scenario [7].

Li et al. (2016) mainly researched the evacuation in the case of fire. The Pathfinder simulation software based on Agent technology simulated the evacuation process of personnel [8]. Hong et al. (2016) simulated five typical evacuation behaviors of subway passengers in emergency situations based on insights from passenger behavior surveys. A comprehensive model for simulating emergency evacuation is proposed. This model is a thorough psychological analysis of subway passengers and combines five typical behavior models in different scenarios [9]. Zong et al. (2017) proposed a hybrid evacuation model considering pedestrian evacuation time and density to formulate the optimal evacuation plan, and proposed a co-evolving multi-particle swarm optimization method to simulate the evacuation process of pedestrians and vehicles and Interaction between these two modes of transportation [10]. Zhang et al. (2019) studied the length of the evacuation route, the time of the evacuation process, and the pedestrian flow density during the peak hours of the subway. A multi-attribute decision-making method based on system simulation, which takes into account the complex interaction between pedestrians and traffic, and models the uncertainty and dynamics of pedestrian behavior during evacuation [11].

Lo et al. (2004) introduced the evacuation model-Spatial Grid Evacuation Model (SGEM), which includes a pre-processing engine to assist in converting spatial information from computer-aided design (CAD) -based architectural plans and performing simulations in many complex buildings, escape patterns are generated in the substance [12]. Sheffi (1982) introduced NETVACl, a model that simulates traffic patterns during emergency evacuation [13]. Shi (2018) proposed a new grid-based mesoscopic model of evacuation dynamics, which was used to calculate the dynamically changing population density in the metacell during evacuation, and the simulation software AnyLogic was used to evaluate the model [14]. Kakizaki (2017) discusses three-dimensional (3D) mass most sparse simulations using precise motion digital human (KDH) models and experimental studies [15]. Abroad research on subway evacuation models is rich and comprehensive, almost involving all aspects of the efficiency, strategy and simulation of subway evacuation [16–18]. Simulation is a more intuitive way to study the situation of subway evacuation. It is an indispensable method in the research of subway evacuation. People are the protagonists of subway evacuation [19, 20]. The most important thing in subway evacuation is the safety of people. About people, in the subway evacuation process, changes in psychology, behavior, etc. are the key to model research; the overall evaluation of subway evacuation is a prerequisite for subway safety, and the cellular automata evacuation model has been a very hot research direction in recent years, it can be used as an extended research in this paper.

3 Platform Evacuation Algorithm

In a densely populated space, when a disaster occurs, “evacuation” is undoubtedly the main means to avoid more serious disasters. No matter from the perspective of safety management or the self-rescue of people in disasters, evacuation will be taken consciously.

Pathfinder mainly includes two aspects: one is the setting of the evacuation environment, and the other is the behavior setting of the evacuees. Through two aspects of settings, the evacuation path is judged by the multi-agent, and then the evacuation time and conditions that may occur during the evacuation, such as congestion, are calculated. Pathfinder belongs to a category of BIM (Building Information Modeling) software. However, Pathfinder is not similar to 3DsMax, Revit and other software that tends to architectural design and architectural performance. Pathfinder’s evacuation environment information settings serve the crowd’s evacuation behavior, which focuses on the evacuation area, evacuation exit, floor access, obstacles, etc., are not dedicated to the concrete performance of the evacuation environment. Pathfinder’s other main content is to set the behavior of evacuees, including the type of evacuation speed distribution of evacuation groups (such as normal distribution, uniform and distribution, constant, etc.), shoulder width, whether to use equipment (such as wheelchairs, etc.), escape exit Select etc. Through the cooperation of two functions, the evacuation route and time of the simulated crowd are calculated.

Pathfinder software simulates crowd evacuation scenarios very well, and provides a good tool for the verification, experiments, and simulation of crowd evacuation theory. It has a very good application in academic or fire protection practice and teaching.

In Pathfinder, the evacuation mode uses the steering mode, because the steering mode is closer to the emergency evacuation behavior of the crowd. The movement of multi-agents in Steering mode conforms to the following algorithm:

(1) Expected speed and acceleration at the starting position:

$$V_1 = V_{\max} \times \frac{0.85k}{1.19} \tag{1}$$

$$a_{\max} = \frac{V_{\max}}{t_{\text{accel}}} \tag{2}$$

Among them, V_1 indicates the expected speed of the starting position; $V_{\max}$ indicates the maximum speed of the multi-agent; k indicates the evacuation influence factor, which is determined by the evacuation environment, and the plane and stairs are different; $a_{\max}$ indicates the maximum acceleration; t_{accel} represents the acceleration time, which is set according to the situation.

(2) Determine evacuation direction:

$$C_{\text{seek}} = \frac{\theta_{\text{t}}}{2\pi} \tag{3}$$

Among them, C_{seek} represents the weight of each direction; θ_{t} represents the angle between the optimal evacuation curve seek curve and the speed direction.

(3) Expected speed and acceleration when moving:

$$|\overrightarrow{v_1}| = \begin{cases} 0 \ (d_{\max} \le d_{\text{stop}}) \\ v_1 (d_{\max} > d_{\text{stop}}) \end{cases} \tag{4}$$

$$\overrightarrow{v_2} = |\overrightarrow{v_1}| \overrightarrow{d_1} \tag{5}$$

$$|\overrightarrow{a_1}| = \frac{\overrightarrow{v_2} - \overrightarrow{v_1}}{|\overrightarrow{v_2} - \overrightarrow{v_1}|} a_{\max} \tag{6}$$

Among them, $\overrightarrow{V_1}$ is the vector speed of the current moving direction; $d_{\max}$ is the maximum walking distance in this direction; d_{stop} is the shortest acceleration distance in this direction; $\overrightarrow{V_2}$ is the speed vector in the minimum moving direction of the weight; $\overrightarrow{d_1}$ is the direction of the evacuation vector in this direction; $\overrightarrow{a_1}$ is the vector acceleration in direction.

(4) Vector formula of speed and position to reach location:

$$\overrightarrow{v_3} = \overrightarrow{v_2} + \overrightarrow{a_1} \Delta t \tag{7}$$

$$\overrightarrow{q_2} = \overrightarrow{q_1} + \overrightarrow{v_2} \Delta t \tag{8}$$

where $\overrightarrow{V_3}$ is the vector velocity at the arrival point;$\overrightarrow{q_1}$, $\overrightarrow{q_2}$ are the arrival point and the current position respectively; Δ_t is the interval time.

4 Evacuation Modeling Simulation

4.1 Simulation Parameter Design

(1) *Number of people*

The “Metro Safety Evacuation Code” issued by the China National Standardization Administration Committee will calculate the average movement speed of people according to the following formula, taking into account the gender and age ratio of

Table 1 Pedestrian evacuation behavior data Table 1

Different people	Horizontal walking speed (m/s)	Downstairs speed (m/s)	Stair up speed (m/s)
Young men	1.25	0.9	0.67
Young women	1.05	0.74	0.63
Olds and kids	0.76	0.52	0.4

the evacuees, and performing arithmetic average:

$$V = \sum_i C_i V_i \tag{9}$$

V represents the average movement speed of the personnel; V_i indicates the speed of different groups of people, and gives a statistical table of the speed of different groups; C_i indicates the proportion of the three types of people divided by age and gender (Table 1).

The simulation data is set in the existing trusted data, and reasonable adjustments are made according to the model scenario and fire situation.

(2) *Crowd density*

The flow of evacuated people has an important relationship with the density of evacuated people. Crowd density is a quantitative indication of the degree of crowd density, the number of people on a unit area is generally people/m^2. Fruin [20] proposed a standard for pedestrian flow service, the more comfortable traffic density is 0.55–2.15 people/(m^2 s)in the horizontal direction, the walking stairs up is 0.33–0.87 people/(m^2 s), the stairs downstairs is 0.44–1.27 people/(m^2 s). During evacuation, the speed of crowd movement depends mainly on the density of the crowd, and not on the average movement speed of the individuals in the crowd. There are even scholars who give empirical diagrams of crowding and crowd movement speed.

4.2 *Platform Evacuation Modeling Simulation*

The platform is the space where passengers and subway cars connect. When the platform fire, according to the evacuation requirements, drive into the subway cars that have not stopped at the platform, instead of stopping at the fire station, go directly to the next station. Therefore, in the event of a fire, only the evacuation of the crowd at the station can be considered.

This section uses island platforms as an example to discuss the simulation of crowd evacuation after a platform fire.

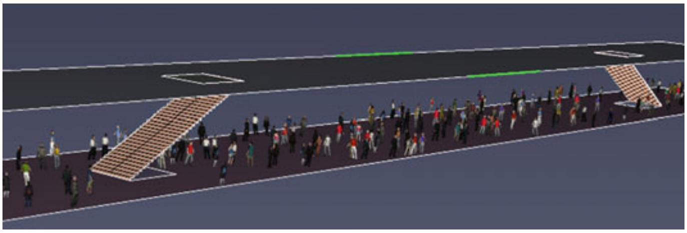


Fig. 1 Platform evacuation model

(1) *Model description*

In this section, based on the established FDS, an island-type platform evacuation Pathfinder model is established, the platform model is divided into two layers, the first layer is the platform, the planar area is 120 m × 14 m and the height is 6 m; the area of second is same with the first; two pedestrian ladders are connected between the first and second floors, and the width of the stairs is 4 m. Pedestrians are evacuated from the platform on the first floor through the pedestrian ladder to the safety exit between the two sides of the second floor, the safety exit is 10 m long. The first floor randomly distributed 200, 400, and 600 evacuation pedestrians.

The model is shown by Fig. 1:

(2) *Simulation results*

(a) *Evacuation scenario 1*

It can be seen from the simulation: at the 10th second, the evacuation personnel have left the middle area of the first floor; at the 40th second, the evacuation personnel have left the first floor platform; at the 55th second, the evacuation personnel have left the first and second floors Walking ladder between floors; at 75 s, pedestrians were basically evacuated.

The simulation results showed that the personnel were evacuated in an orderly manner and no congestion occurred. The evacuation route is shown in Fig. 2:

The evacuation route shows that after pedestrians enter the first floor through the walking ladder, the evacuation route is concentrated and orderly, indicating that the evacuation situation is good after the evacuees enter the first floor.

The remaining numbers of the first and second floors are shown in Fig. 3:

The number of people on the first floor was completely evacuated at 42 s; at the 15th second, evacuated personnel began to appear on the second floor, reaching a peak at 49 s, 122 people were evacuated, at the 80th second, the evacuation of personnel on the second floor was completed.

During the evacuation process, the walking ladder became a bottleneck in the evacuation facility, so the number of people on the two walking ladders was counted. The statistics are shown in the following Fig. 4.

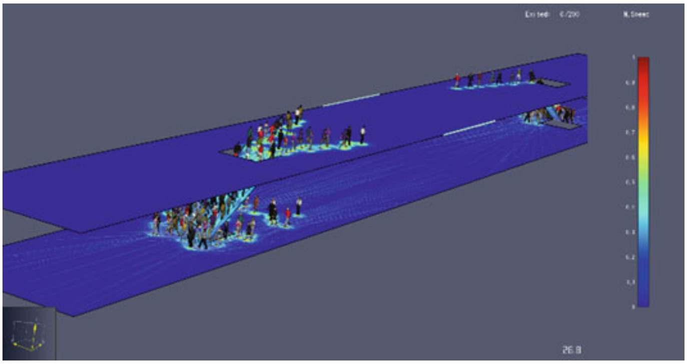


Fig. 2 Evacuation situation 1 at the station evacuation route

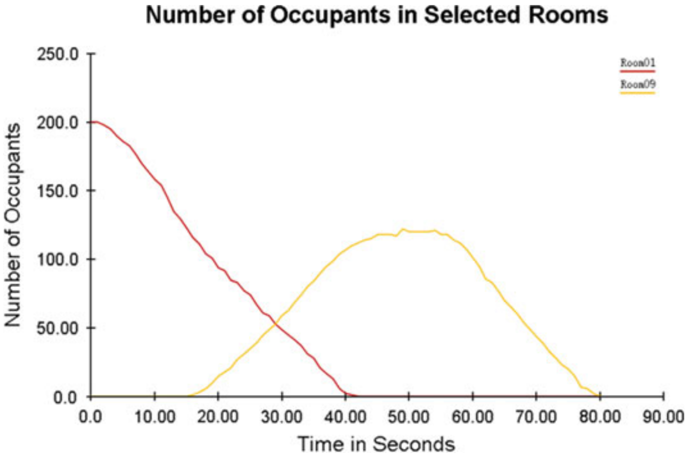


Fig. 3 Statistics of the number of people on the first and second floors of the simulation scenario 1

It can be seen from the statistical figure that the peaks of the two stairs are not same, stair01 enters the peak period from 18 to 38 s, and the peak number is 40–45 people/s; Stair02 reaches the peak in 30–32 s and is 59 people. This is due to the fact that the initial position of the people on the negative platform is not randomly distributed, and the number of pedestrians near Stair01 is relatively small.

(b) *Evacuation scenario 2*

It can be seen from the simulation results that although the number of people has increased to 400, the evacuation of the crowd is still relatively orderly, and no large crowds occurred. At the 50th second, the crowd has basically left the platform; at the 80th second, the crowd has basically left the two-story connected ladders, and the left-hand ladder left earlier; at the 100 s, only 13 people have not left the second-floor platform.

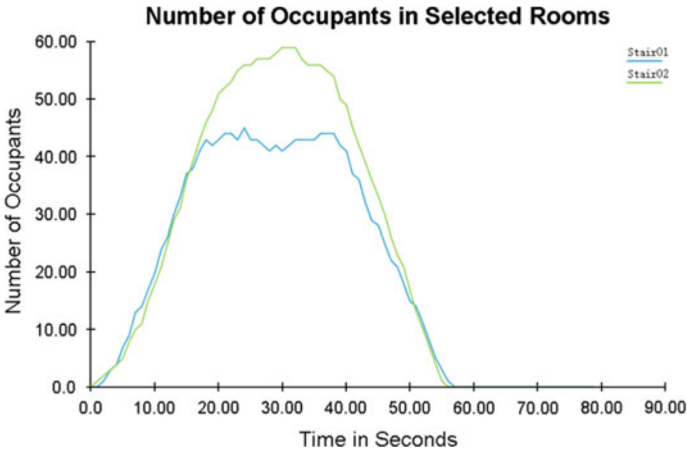


Fig. 4 Statistics of the number of pedestrian ladders

The evacuation route figure is as follows Fig. 5. The evacuation route shows that although there are a large number of evacuees, there is no serious congestion and the evacuation is orderly.

Evacuation simulation scenario 2, the remaining number of the first floor and the second floor is shown in Fig. 6:

The statistical figure shows that at the 16th second, evacuated crowds began to appear on the second floor; at the 78th second, the number of people on the second floor reached a peak of 167; the number of people on the first floor dropped evenly after the third second.

The number of evacuated people on the walking ladder is shown in the following Fig. 7:

As can be seen from the statistical figure, Stair01 reached a peak at 48 and 51 s, respectively, for 92 people; Stair02 reached a peak at 54 s, for 90 people.

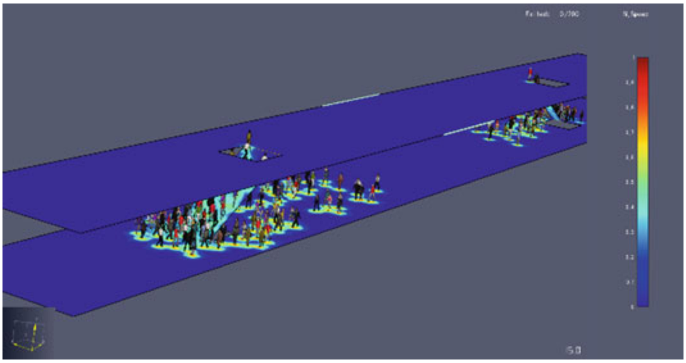


Fig. 5 Evacuation situation at the station 1 Evacuation route

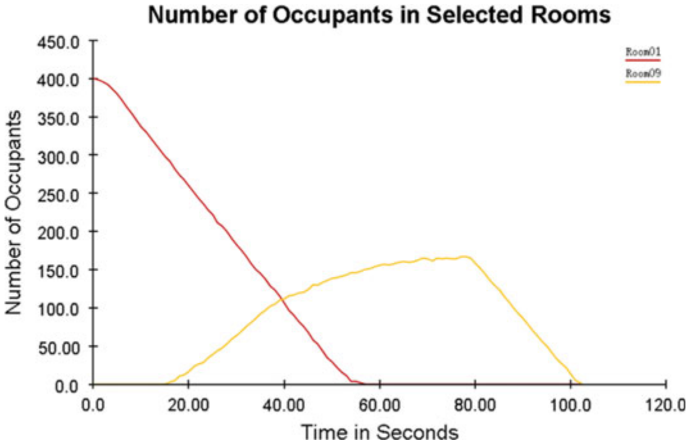


Fig. 6 Simulation scenario 2 of platform evacuation 1st and 2nd floor statistics

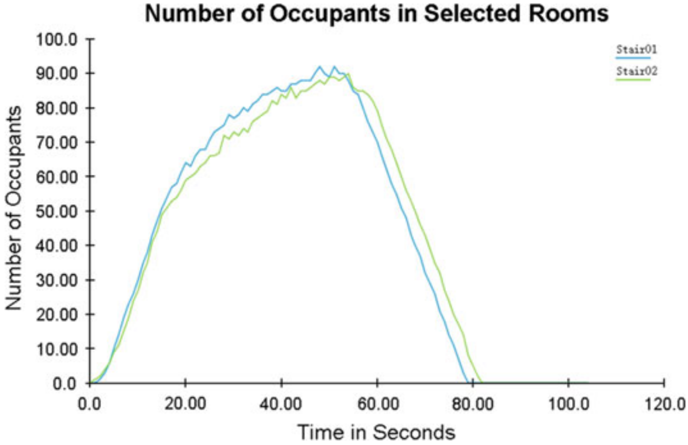


Fig. 7 Statistics of the number of pedestrian ladders

(c) *Evacuation scenario 3*

The simulation results show that at the 15th second, the evacuated crowd gathered toward the pedestrian ladder, and a certain gathering range was formed, and the congestion was obvious; at the 83th second, the crowd was evacuated from the first floor platform; at the 107th second, the evacuation crowd left the ladder; at the 129th second, the crowd was evacuated.

In the 30–60 s time interval, the pedestrian ladders gathered more. The heat map of the evacuation crowd density at 30 s is shown in Fig. 8:

It can be seen from the pedestrian evacuation route map that although the evacuation is more orderly, there are more crowds at the pedestrian ladder, and the escape

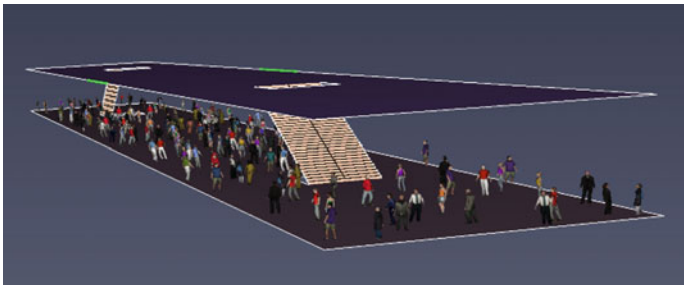


Fig. 8 The heat map of the density of people evacuated from the station at $T = 30$ s

route at the pedestrian ladder is chaotic. However, no aggregation phenomenon was formed at the evacuation exit on the second floor (Fig. 9).

Figure 10 shows the statistics of the number of people on the first and second planes: after 3 s, the number of people on the first floor drops uniformly; on the second floor, evacuation begins to occur at 16th second, and the peak occurs at 105th

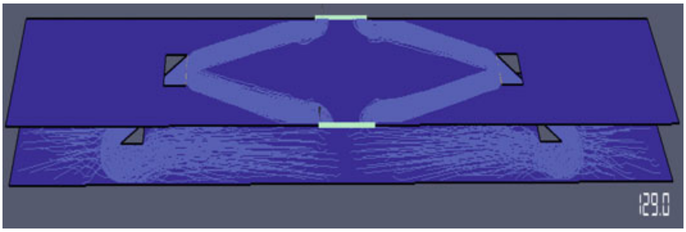


Fig. 9 Pedestrian evacuation route

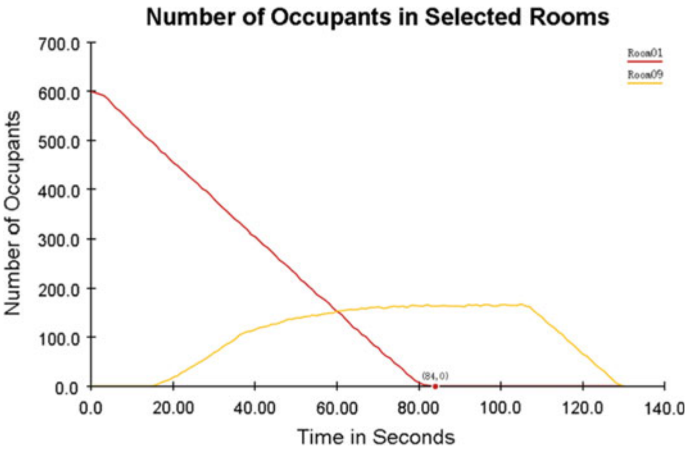


Fig. 10 Statistics of the number of people on the first and second floor planes

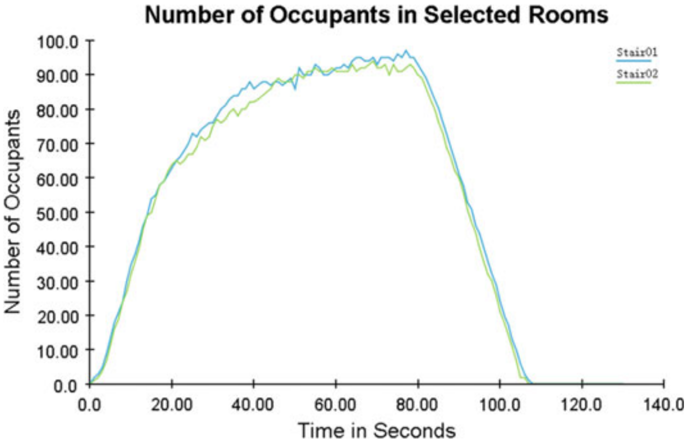


Fig. 11 Number of evacuated people on the left and right stairs

second, reaching 168 people; On the 129th second, the evacuation of the personnel on the second floor was completed.

The number of stairs on the left and right is as shown in the Fig. 11. After 15 s, the number of stairs on both sides of the stairs began to slow down, reaching peaks at 77 and 78 s, 97 and 93 people respectively. After that, the number of people on the stairs began to drop evenly.

(3) *Analysis of simulation results*

Comparing station simulation scenario 1, scenario 2 and scenario 3, although the number of persons in scenario 2 is twice that of scenario 1, it did not cause serious congestion and the evacuation proceeded in an orderly manner; the number of evacuation in scenario 3 increased to 600 people at the entrance of the pedestrian ladder Congestion has formed, causing delays in evacuation time. Evacuation was completed in 80 s in scenario 1, evacuated in 102 s in scenario 2, and 129 s in scenario 3. From 200 to 400 people, it took 22 s to evacuate; from 400 to 600 people, it took 27 s to evacuate. In scenario 1, the number of people on the second floor plane is in the shape of a regular figure, in scenarios 2 and 3, the left side of the trapezoid is stretched, which indicates that as the number of simulations increases, the time for the crowd to enter the second floor plane is lengthened, the evacuation time from the peak to the complete evacuation of the crowd was almost unchanged, all around 20 s.

5 Conclusion

In this paper, simulations were carried out for station platforms, firefighter walking ladder retrograde, and different distributions of crowd density. By simulating the above scenarios and adjusting the simulation parameters respectively, the crowd evacuation scenarios under different conditions and parameters were observed. Comparison of simulation results for different scenarios. Key data such as crowd velocity during evacuation, number of people passing through space and remaining, number of people passing through exits, evacuation paths, etc. were Record. Testing the crowd evacuation in different situations. From the above simulation test results, we can see that crowd flow, density and station structure have a direct impact on the evacuation effect.. Through the above analysis, suggestions for optimizing the evacuation of rail transit platforms include two aspects: on the one hand, the evacuation process should keep the crowd orderly evacuation to avoid panic; on the other hand, platform access should be designed with full consideration of evacuation requirements, avoiding narrow exits, and Guidance signage should be clear and unambiguous.

There are still some shortcomings in the findings of this paper, which do not take into account the gender and age of the passengers, as well as the lack of a clear understanding of passengers' The analysis of the influence of psychological characteristics on passenger behavior. Therefore, in the subsequent simulation study, passengers were categorized into male and female by gender, and older, younger, and younger by age. Children, with due consideration given to the changing psychological profile of passengers in sudden-onset disaster situations and the implications for evacuation behavior.

References

1. J.C. Ma, X.Y. Zhou, J. Li, Application of computer simulation method to parameter influence study of crowd evacuation. J. Natural Disasters **18**(6), 154–159 (2009)
2. Dai, B., Wang, T., Qin, Y., Subway station evacuation simulation system, in *International Conference on Measuring Technology & Mechatronics Automation* (2009)
3. Kang, K., A stochastic evacuation model for fire life safety assessment in transportation system, in *Pedestrian and Evacuation Dynamics* (Springer, US, 2011)
4. X. Han, J.N. Ma, B.H. Cong, Simulation analysis on crowd evacuation of the subway train fire. Adv. Mater. Res. **424–425**, 1215–1219 (2012)
5. Yang, P., Li, C., Chen, D., *Fire Emergency Evacuation Simulation Based on Integrated Fire–Evacuation Model with Discrete Design Method* (Elsevier Science Ltd., 2013)
6. D.F. Zhao, X.L. Zhang, Y. Su, Reliability analysis of relevant radom variables in pedestrian evacuation study in underground transportation hub. Adv. Mater. Res. **1020**, 741–746 (2014)
7. Y. Cai, Z.Y. Lin, J. Mao et al., Study on law of personnel evacuation in deep buried metro station based on the characteristics of fire smoke spreading. Proc. Eng. **135**, 543–549 (2016)
8. Z.Y. Li, M. Tang, D. Liang et al., Numerical simulation of evacuation in a subway station. Proc. Eng. **135**, 616–621 (2016)
9. L. Hong, J. Gao, W. Zhu, Simulating emergency evacuation at metro stations: an approach based on thorough psychological analysis. Transp. Lett. **8**(2), 113–120 (2016)
10. Zong, X., Wang, C., Chen, H., An evacuation model based on co-evolutionary multi-particle swarms optimization for pedestrian–vehicle mixed traffic flow. Int. J. Modern Phys. C 1750142 (2017)
11. L. Zhang, M. Liu, Wu. Xianguo et al., Simulation-based route planning for pedestrian evacuation in metro stations: a case study. Autom. Constr. **71**, 430–442 (2016)

12. S.M. Lo, Z. Fang, P. Lin et al., An evacuation model: the SGEM package. Fire Saf. J. **39**(3), 169–190 (2004)
13. Sheffi, Y., A transportation network evacuation model. Transp. Res. A **16** (1982)
14. M. Shi, E.W.M. Lee, Yi. Ma, A novel grid-based mesoscopic model for evacuation dynamics. Phys. A **497**, 198–210 (2018)
15. T. Kakizaki, J. Urii, M. Endo, Simulation and experiment of mass evacuation to a Tsunami evacuation tower. ASCE-ASME J. Risk Uncertainty Eng. Syst. Part B: Mech. Eng. **3**(4), 041007 (2017)
16. J. Zeng, Q.G. Yao, Y.S. Zhang, J.T. Lu, M. Wang, Optimal path selection for emergency relief supplies after mine disasters. Int. J. Simulation Modelling **18**(3), 476–487 (2019)
17. D. Gong, M. Tang, S. Liu, G. Xue, L. Wang, Achieving sustainable transport through resource scheduling: a case study for electric vehicle charging stations. Adv. Prod. Eng. Manage. **14**(1), 65–79 (2019)
18. T. Gocken, A.T. Dosdogru, A. Boru, M. Gocken, Integrating process plan and part routing using optimization via simulation approach. Int. J. Simulation Modelling **18**(2), 254–266 (2019)
19. U. Zuperl, F. Cus, A cyber-physical system for smart fixture monitoring via clamping simulation. Int. J. Simulation Modelling **18**(1), 112–124 (2019)
20. J. Janekova, J. Fabianova, M. Fabian, Assessment of economic efficiency and risk of the project using simulation. Int. J. Simulation Modelling **18**(2), 242–253 (2019)

Optimization of Non-motorized Traffic Around Rail Transit Stations

Ziyao Zhao, Li Wang, and Xiaoning Zhu

Abstract This paper aims to develop the safety, systematic and characteristics of the non-motorized traffic around rail transit stations. A questionnaire is designed and distributed to five stations of Nanchang rail transit. After statistics and reliability analysis of the collected sample data, the structural equation model (SEM) is established. The results of the model are analyzed in detail, and suggestions for the next step of non-motorized optimization are put forward. Besides, this paper plans the non-motorized traffic around the rail transit station from three levels, and takes Bayi Square as an example to show the specific micro road design.

Keywords Non-motorized traffic · Rail transit stations · Optimization

1 Introduction

As an important part of urban transportation, urban rail transit is developing rapidly in China. However, with the rapid development of urbanization, road traffic flow is becoming more and more saturated, and traffic problems are becoming increasingly serious.

Compared with motorized transportation, non-motorized transportation is a green and environmentally friendly way of travel. The development of non-motorized transportation helps to reduce energy consumption, traffic congestion and environmental pollution, and promote the sustainable development of the city.

Guided by the direction of sustainable urban development, the non-motorized traffic optimization has gradually attracted the attention of researchers. It is an

Z. Zhao (✉) · L. Wang · X. Zhu
Beijing Jiaotong University, Beijing, China
e-mail: 19125796@bjtu.edu.cn

L. Wang
e-mail: liwang@bjtu.edu.cn

X. Zhu
e-mail: xnzhu@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_16

inevitable trend of urban traffic to develop non-motorized traffic according to local conditions.

This paper is organized as follows. In Sect. 2, a brief review of previous works is given. Factor analysis based on structural equation model is introduced in Sect. 3. Non-motorized optimization is proposed in Sect. 4. Section 5 gives conclusions.

2 Literature Review

Parsons Brinckerhoff selected different rail transit stations for field investigation, and proposed improvement measures for pedestrian facilities [1]. Kevin and Eric conducted a cost–benefit assessment to study travelers' preference for connecting facilities [2]. Bernhard Snizek revised the bicycle lane by presupposition factors. [3]. Dickins selected two representative urban studies to investigate the utilization rate of non-motorized-moving space facilities around the site, and concluded the factors affecting the layout of facilities [4]. Chenyin Long based on the use demand of non-motorized space around urban rail transit stations, taking the non-motorized space around Chongqing rail transit stations as an example, put forward a specific construction plan [5]. Yanni Weng enumerated the existing problems and put forward reasonable treatment methods for Beijing rail transit stations [6].

The above studies provide an important theoretical basis for non-motorized traffic optimization. However, there are still some aspects of non-motorized traffic optimization that are not fully discussed.

3 Influencing Factors Analyzing Based on Structural Equation

3.1 Questionnaire Design

The main body of the questionnaire is divided into two parts. The first part is the individual travel information survey, which mainly obtains some personal information of citizens. The second part is their evaluation of traffic environmental impact factors around the station.

3.2 Reliability Test of Questionnaire

A total of 521 valid questionnaires are collected. The questionnaire questions were analyzed by SPSS, and Cronbach α was 0.851, which was greater than 0.8. This value shows that the research data has high reliability quality.

3.3 Structural Equation Model (SEM) Analysis

SEM consists of measurement model and structural model. SEM can not only study the relationship between the obtained data, but also find the relationship between the set "potential variables".

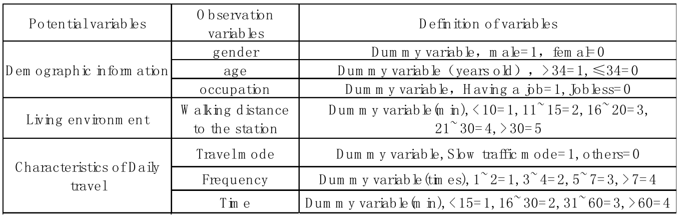
(1) *Variable description* (Table 1)

(2) *Model establishment and result analysis* (Fig. 1; Table 2)

The path index of the satisfaction of the aspects to slow traffic facilities are 0.312, 0.314, 0.358, 0.324 and that of satisfaction to daily travel characteristics is 0.728, which all pass the significance test. The fitting results show that: There is a certain degree of difference in the satisfaction of different groups; different groups have differences in the choice of daily travel mode; different living environment will have an impact on the satisfaction of non-motorized facilities and daily travel mode. From the path coefficient between potential variables, we can know that the total path coefficient of satisfaction of non-motorized facilities on daily travel characteristics is 0.835, of which the path coefficient of direct impact is 0.728, and the path coefficient of indirect impact through living environment is 0.227, which shows that the indirect impact of non-motorized facilities satisfaction through the intermediary variable of living environment is less than the direct impact of non-motorized facilities satisfaction.

Among the nine indicators reflecting the satisfaction degree of non-motorized traffic facilities, the influence degree of standardized path coefficient on the satisfaction degree of non-motorized traffic facilities is as follows: non-motorized traffic conditions around the station (0.732), green facilities (0.662), parking lot (0.629), public transport station (0.624), bicycle placement point (0.61), guide sign (0.597), lighting facilities (0.596) and none Barrier (0.59). Among them, the standardized path coefficient of non-motorized road condition is 0.695, which has the greatest impact on the satisfaction of the citizens. Therefore, the non-motorized road condition around the station is the most important part of the non-motorized space optimization around the station. This result will be used especially in micro optimization.

Table 1 Variable description



Potential variables	Observation variables	Definition of variables
Demographic information	gender	Dummy variable，male=1，femal=0
	age	Dummy variable（years old），>34=1, ≤34=0
	occupation	Dummy variable，Having a job=1, Jobless=0
Living environment	Walking distance to the station	Dummy variable(min), <10=1, 11~15=2, 16~20=3, 21~30=4, >30=5
Characteristics of Daily travel	Travel mode	Dummy variable, Slow traffic mode=1, others=0
	Frequency	Dummy variable(times), 1~2=1, 3~4=2, 5~7=3, >7=4
	Time	Dummy variable(min), <15=1, 16~30=2, 31~60=3, >60=4

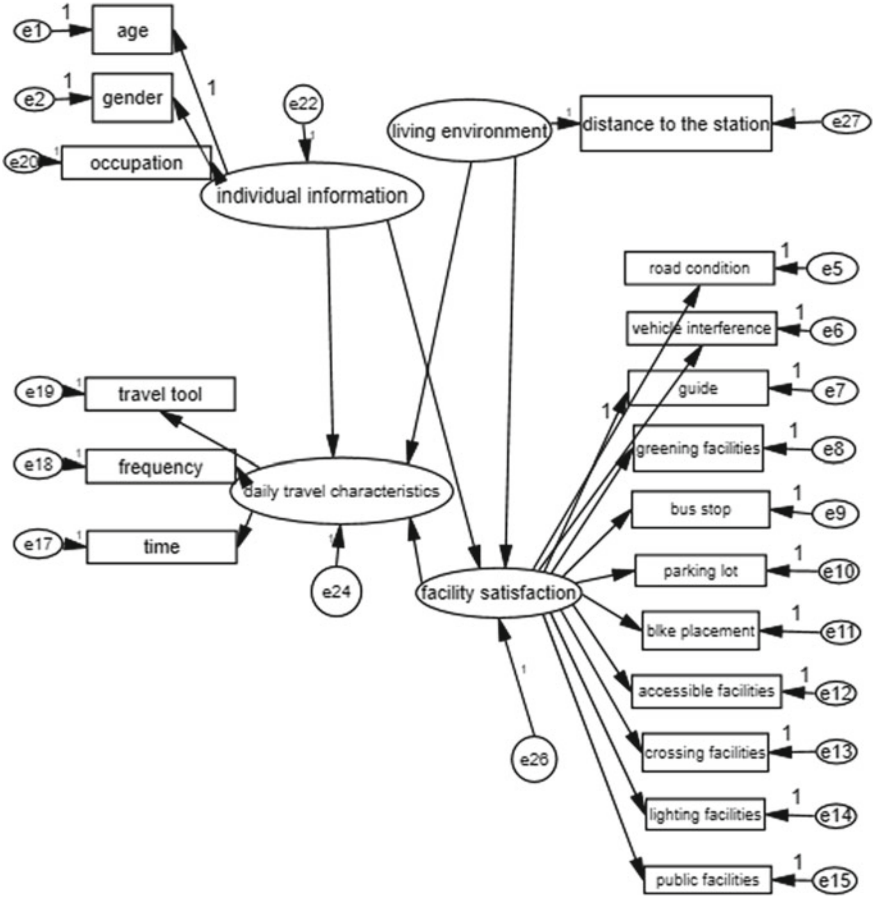
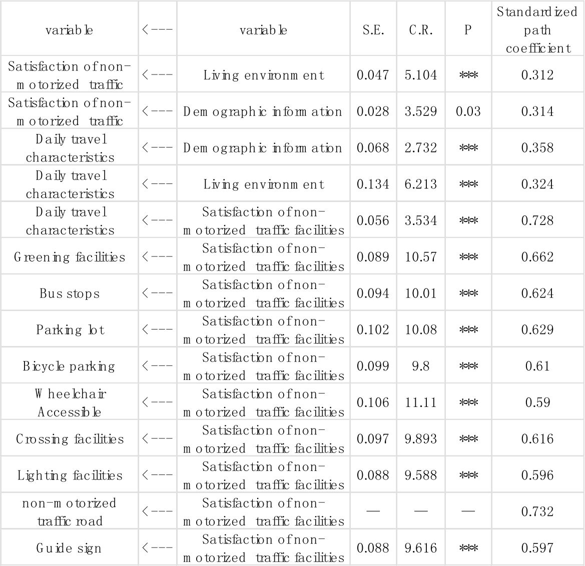


Fig. 1 Schematic diagram of SEM

4 Non-motorized Traffic Optimization Around Rail Transit Station

In this paper, the optimization of non-motorized traffic around rail transit stations is summarized as four principles: systematic, people-oriented, diversity and characteristics.

Table 2 Estimation results of path coefficient of structural equation



variable	<---	variable	S.E.	C.R.	P	Standardized path coefficient
Satisfaction of non-motorized traffic	<---	Living environment	0.047	5.104	***	0.312
Satisfaction of non-motorized traffic	<---	Demographic information	0.028	3.529	0.03	0.314
Daily travel characteristics	<---	Demographic information	0.068	2.732	***	0.358
Daily travel characteristics	<---	Living environment	0.134	6.213	***	0.324
Daily travel characteristics	<---	Satisfaction of non-motorized traffic facilities	0.056	3.534	***	0.728
Greening facilities	<---	Satisfaction of non-motorized traffic facilities	0.089	10.57	***	0.662
Bus stops	<---	Satisfaction of non-motorized traffic facilities	0.094	10.01	***	0.624
Parking lot	<---	Satisfaction of non-motorized traffic facilities	0.102	10.08	***	0.629
Bicycle parking	<---	Satisfaction of non-motorized traffic facilities	0.099	9.8	***	0.61
Wheelchair Accessible	<---	Satisfaction of non-motorized traffic facilities	0.106	11.11	***	0.59
Crossing facilities	<---	Satisfaction of non-motorized traffic facilities	0.097	9.893	***	0.616
Lighting facilities	<---	Satisfaction of non-motorized traffic facilities	0.088	9.588	***	0.596
non-motorized traffic road	<---	Satisfaction of non-motorized traffic facilities	—	—	—	0.732
Guide sign	<---	Satisfaction of non-motorized traffic facilities	0.088	9.616	***	0.597

4.1 Overall Non-Motorized Space Layout

(3) *Set up radial ring road network centered on stations*:
The setting of the ring road can also effectively divert the motor vehicle travel through the station, reduce the traffic pressure, reduce the conflict between non-motorized traffic and motor vehicle traffic, and improve the safety and comfort of non-motorized travel. Radial roads can be combined with urban roads, and can also be used as pedestrian corridors of urban buildings or park greenways (Fig. 2).

(4) *Ensure good connectivity of blocks around the station*:
According to the passenger flow direction of the rail station, pedestrian corridor is added to ensure the good connectivity. The commonly used methods are: (a) add small-scale branches to connect the station and the surrounding road network. (b) open a pedestrian space from the inside of the building in the street area (Fig. 3).

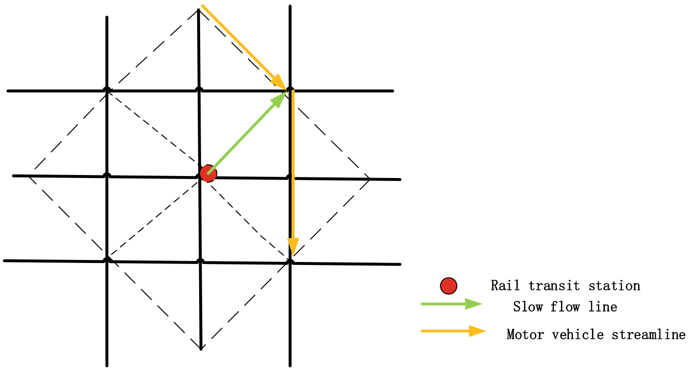


Fig. 2 Schematic diagram of radial ring road network structure

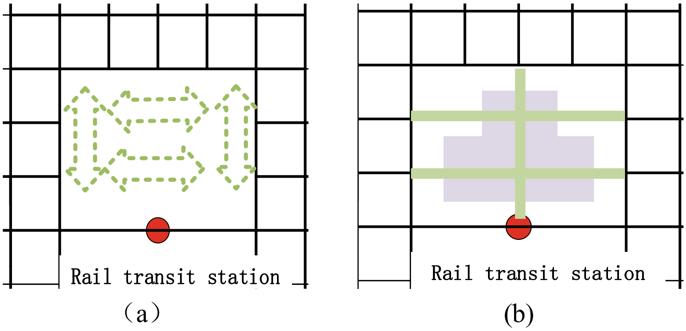


Fig. 3 Schematic diagram of opening large blocks around the station

(5) *Building a three-dimensional walking network:*
A three-dimensional walking network is required to establish a continuous, efficient and comfortable traffic connection between different levels of rail stations (Fig. 4).

4.2 Optimize the Key Nodes of Non-Motorized Traffic

(1) *Public transport connection*:

(a) *Sign guide for public transport connection:* The identification guidance of public transport transfer should start from the inside of the station hall, and make the transfer crowd have a clear understanding of their location.

(b) *Optimal layout of public transportation lines:* Explore the development of new public transport service forms, optimize the public transport dispatching

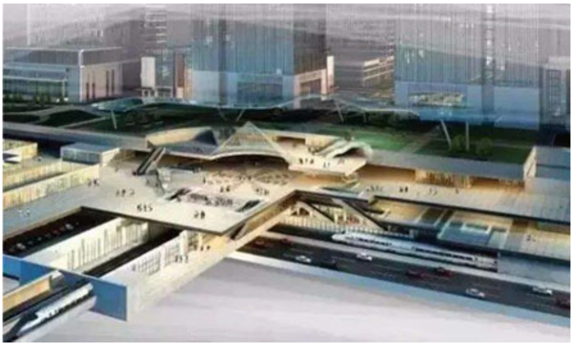


Fig. 4 Construction form of three-dimensional pedestrian network around the station

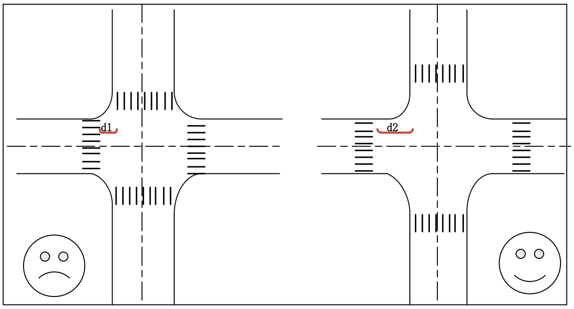


Fig. 5 Setting up of crosswalk at intersection

management and operation plan in peak period, and create more favorable conditions for the transfer between public transport and rail transit.

(2) *Optimization of non-motorized crossing system*:
Figure 5 shows the setting of crosswalks at intersections. In order to avoid pedestrians meeting at intersections in adjacent directions, the setting of crosswalks needs to be a distance from the corner. In addition, it is also convenient for right-turning vehicles to notice the pedestrians at the corner and improve safety.

4.3 Design of Humanized Non-motorized Traffic Road Facilities:

From the results of structural equation analysis in Chap. 2, We should pay more attention to the optimization of humanized non-motorized traffic road facilities. This paper carries out micro optimization of non-motorized traffic road facilities from three aspects: Lane division, pedestrian flow relief and bicycle lane improvement.

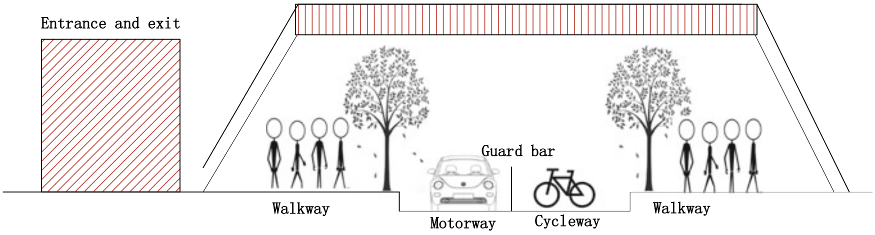


Fig. 6 Modification of cross section of Bayi Square Station after entrance and exit

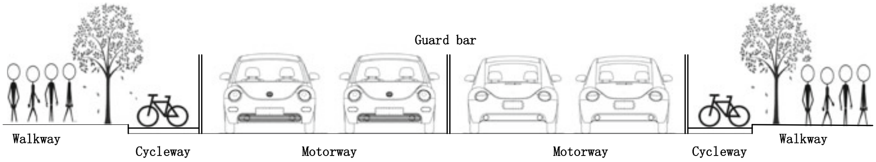


Fig. 7 Section picture of lane reconstructed

(1) *Site Flow Dispersion*
The entrance and exit of Bayi Square Station is a place where people gather. For the convenience of evacuation and the construction of passengers' stay, pedestrian overpasses can be set up to alleviate congestion (Fig. 6).

(2) *Bicycle Lane Safety*:
Raise and isolate the bicycle lane. Raise the bicycle lane level from the motor vehicle level by about 6 cm. Spatially, bicycle lanes are separated from motor lanes. Moreover, due to the existence of low altitude, motor vehicle drivers usually keep a certain distance from the edge of the bicycle lane, thus playing a certain role in protecting the bicycle lane users (Fig. 7).

5 Conclusion

This paper combines the optimization methods and theories of non-motorized-moving space at home and abroad. In this paper, a questionnaire is designed and the structural equation model is established. The results of the model are analyzed, and suggestions for the next step of non-motorized optimization are put forward. Besides, this paper plans the non-motorized traffic around the rail transit station from three levels and takes Bayi Square as an example to show the specific micro road design. Thus, such reasonable optimization of non-motorized traffic around rail transit stations can improve residents' travel happiness and service level.

Acknowledgements This paper was funded by the National Key R&D Program of China (grant number: 2018YFB1201403).

References

1. Brinckerhoff, P., *Metrorail Bicycle & Pedestrian Access Improvements Study*. Washington Metropolitan Area Transit Authority: Washington, DC, USA, 2010
2. K. Kevin, S. Eric, Assessing options to enhance bicycle and transit integration. Transp. Res. Rec. **2217**(1), 162–167 (2011)
3. B. Snizek, Mapping bicyclists' experiences in Copenhagen. J. Transp. Geogr. **30**, 227–233 (2013)
4. D.S. Vale, Transit-oriented development, integration of land use and transport, and pedestrian accessibility: combining node-place model with pedestrian shed ratio to evaluate and classify station areas in Lisbon. J. Transp. Geogr. **45**, 70–80 (2015)
5. Long, C., Investigation and Optimization Strategy of Walking Space Around Chongqing Central Track Station (Chongqing University, 2015)
6. Weng, Y., Yang, L., Brief analysis of non-motorized transit in rail transit station area. Informatization Constr. (7) (2016)

Layout Planning of Container Piggyback Transport Stations Based on Road-Railway Intermodal Transportation

Tian Xia, Xiaoning Zhu, and Li Wang

Abstract In response to the call of Chinese government, accelerate the development of multimodal transport system. This article aims to the key layout planning problem as study object, takes container piggyback transportation as the background and bases on the idea of hub-and-spoke network. Aiming at minimizing the total cost of network operation, establishes the location model of piggyback transport node based on the hub-and-spoke network, designs the heuristic algorithm and combines the current regional cargo traffic and transportation network status in China and then calculate the transportation network hub location scheme and the total operating cost of the hybrid hub-and-spoke network, compared with the pure-axis spoke network and the shortest-circuit network.

Keywords Intermodal transportation · Container transportation · Hub-spoke network

1 Introduction

Piggyback transportation refers to a mode of transport in which one means of transport is placed on another. This article mainly studies the combination of containers and semi-trailers, and the adoption of roll-on or lift mode, as well as the use of railway special vehicles for long-distance transport, known as TOFC (Trailer on flat car). This mode of transportation can effectively combine the characteristics of large transportation volume and low energy consumption in railway with the characteristics of flexibility and convenience in road transportation, so as to reduce the energy consumption and transportation cost of the whole network.

T. Xia (✉) · X. Zhu · L. Wang
Beijing Jiaotong University, Beijing, China
e-mail: 18120919@bjtu.edu.cn

X. Zhu
e-mail: xnzhu@bjtu.edu.cn

L. Wang
e-mail: liwang@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_17

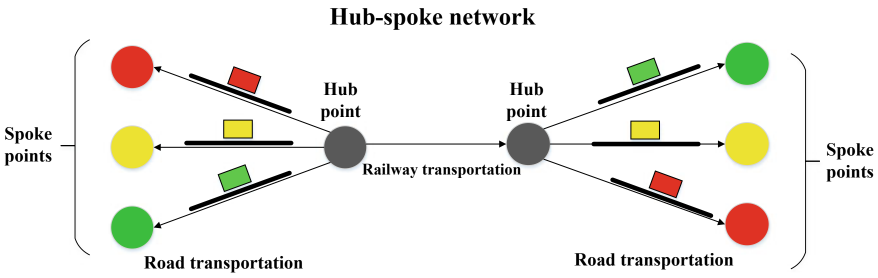


Fig. 1 Hub and spoke piggyback transport network

Piggyback transport has been developed in developed countries for many years, and has formed the "one-stop" transport in the United States and "piggyback transport contractor" transport in Europe. The development of piggyback transport is a systematic project, which needs to involve policies and regulations, highway transport technology, railway transport technology, transport organization, information technology and so on. Since the reform of railway freight transport in China, the implementation of "door-to-door" transport has been intensified, providing great policy support for the development of piggy-back transport in China.

In October 2014, the China State Council issued 'The Medium- and Long-term Plan for the Development of Logistics Industry (2014–2020)', which clearly proposed to accelerate the development of multimodal transport system, and explore the construction of multimodal transport systems such as the multimodal transport represented by the transport of piggy-back, the transport of public rail and the transport of public water represented by the transport of water and roller.

Therefore, this article focus on the key problem of the rail-road container piggyback transportation operation "terminal layout problem", based on the existing railway freight transportation network, scattered all over the country's railway logistics base integrating connection, set up piggy-back form systematic container transportation network. Through integrating hub nodes of network in the supply of goods, and scatter the original rail-road transportation, improve transport efficiency, reduce the total cost of the whole network (Fig. 1).

2 Literature Review

Since 2017, Shenhua group took the piggyback transportation, after that domestic scholars have paid more and more attention to the transportation of piggybacks. Domestic and foreign scholars have been doing research on container piggyback transportation and related fields. The main research include: technical equipment research, economic feasibility and pricing research of transportation, transportation organization research, transportation network research, site layout research, and station operation scheduling research.

2.1 Research on Technical Equipment

Lin Jieliang studied the transportation modes of Container on Flat Car (COFC) in America and Trailers on Flat Car (TOFC) in Europe. Based on the current situation of container transportation in China, he analyzed the feasibility study of hatchback transportation in China [1].

Wu Bilong introduced the development of railway loading and reinforcement methods in China, summarized the development history of European and American railway hatchback transport vehicles, and combined with the railway limit in China, proposed a railway hatchback transport loading and reinforcement scheme adapted to the railway hatchback [2].

2.2 Research on Transportation Economics and Pricing

Professor Morash conducted research on the US freight market and found that US manufacturing companies are more inclined to use the TOFC-COFC method to transport goods [3].

Ning Yu focused on the development of the American railroad piggyback transportation, focusing on the transportation distance of the railway hatchback transport, the traffic volume in the United States and the future development prospects, and believed that piggyback transportation is more competitive if the transport radius is more than 150 km [4].

Theodore Tsekeris conducts research to explore the prediction of network traffic flow, and provides different prediction models for different traffic flow and cargo flow development. It can be used accurately grasp the future freight volume in piggyback transportation [5].

Kai Yang develops a novel modeling framework for the IH&S network design problem to jointly minimize the total transportation cost and maximum travel time requirement in term of critical value [6].

2.3 Research on Transport Organization

Wang Baohua studied the supply, demand and allocation of fast cargo resources and constructed a transportation network evaluation system under the comprehensive transportation system [7].

Lu Xiaofang introduced the difference between the US-European railway piggyback transport line and the transportation organization of the train, and analyzed the changes in the total energy consumption of the US [8].

Waleed Najy consider a novel and uncapacitated hub problem and take flow-dependent economies of scale and congestion into consideration. He introduce benders decomposition to solve multiple-allocation hub-and-spoke network design problem [9].

Sibei Liu design a highway express freight axle and Spoke network with time window, which can be apply in multimodal rail-road transportation system to meet time requirements [10].

2.4 Research on the Transportation Network

Sook Tying Choong studied the problem of empty container allocation under the hub-and-spoke network and how to reduce the total cost associated with moving containers based on meeting container loading requirements [11].

Faisal Alkaabeh used lagranian heuristic and GRASP to optimize the original hub-and-spoke layout model, then he adds some valid inequalities to accelerate the convergence rate of the Lagrangian heuristic [12].

Yan Shangyao studied the design of aviation fast cargo transportation service network, established a multi-commodity network flow model with reference to the hub-and-spoke network, and solved the model by cutting plane method [13].

Ni Linglin analyzed the difference in the operation process between the full connection express network and the hub-and-spoke express network. Through detailed calculation and comparison of the express sorting cost, sorting efficiency, transportation cost and total network cost of the two network modes he concluded that the cost and efficiency of the spoke express network are better than those of the Unicom network [14].

2.5 Research on Station Layout

Weng Kerui introduced the design and algorithms of the hub-and-spoke network. By studying the set coverage problem of multi-distribution hub stations, optimizing the model solving method, and applying the model algorithm results to the optimization problem of China's central hub logistics network [15].

Fan Jun selected the hub-and-spoke transportation station for the national logistics center station, and determined the first and second axis on the basis of site selection, and explained the construction method of constructing the regional logistics channel between the axis [16].

2.6 Review of Research Status

At present, scholars mostly study the equipment selection demonstration and the demonstration stage of the container piggyback transportation, and there are few operations research based on the characteristics of special vehicles for container road hatchback transportation. In the site selection of the station, the problem of cooperation between the container piggyback transport vehicle and the railway station is less considered, and the allocation of the tank source is less considered.

3 Problem Description and Model Formulation

3.1 Network Scale Benefit

The main goal in the layout of the hub-and-spoke network selection node is to use the economies of scale between the hub nodes to minimize the total network cost. In a hybrid hub-and-spoke network, economies of scale are affected by cost sharing, transportation resource utilization, and transportation efficiency. In container piggyback transportation, the use of hub-and-spoke network may bring some road freight vehicle waiting and unnecessary bypass costs, but in the context of the overall scale efficiency of the network, the overall freight cost can still be effectively reduced. In the model construction, using α to represent the scale benefit as a fixed freight discount coefficient ($0 < \alpha < 1$).

3.2 The Number of Hub Nodes and Choose the Layout of the Hub

The location of the hub is based on the size of the node's freight volume and the accessibility of the node to other nodes in the network and optimize layout according to the economic and logistics development of the region. Meanwhile, the appropriate number of hub stations is assessed in conjunction with the actual situation.

Previous research have found that when the number of hub-and-spoke network hubs is $\sqrt{N}$ (where is the total number of nodes in the network), the total cost is lower. Therefore, this article defines the number of railway container piggyback transportation network hubs as $\sqrt{N}$, which is the integer after the square root of the total number of nodes in the network.

3.3 Network Node Transportation Route and Transportation Cost

In this article, by calculating the shortest distance of the transportation distance from any node in the network to each node, and obtaining the route, then the sum of the shortest distances of each node in the network can be obtained. In this hub-spoke

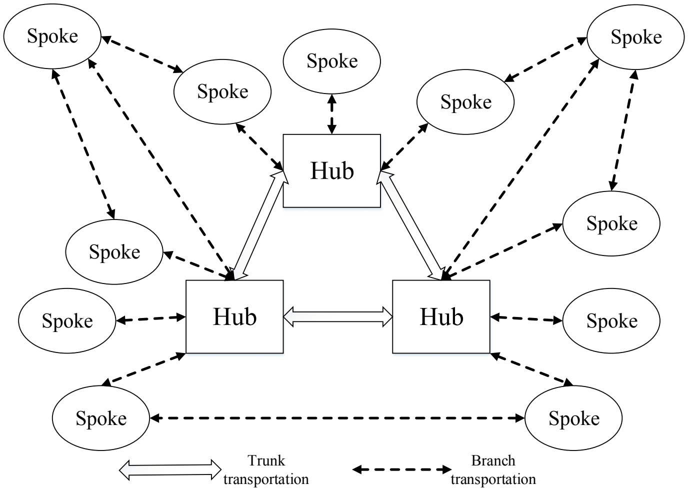


Fig. 2 Multi point mixed hub and spoke network

transportation, the shortest route will also be used for the transport route between the hub points. If the hub transit cost between the hub points is higher than a certain threshold, the hub-spoke transport will be abandoned, and using the shortest route for direct transport. If the transfer between the two points, from the starting point to the end point, will be carried out by means of a spoke-hub-spoke transport, so that the transport route would pass through one or more hub stations (Fig. 2).

This article mainly focuses on the problem of selecting points in the multi-distribution hub-spoke network. In the research process, only the core hub nodes in the multi-distribution network are studied, considering the assignment of the first-level center and the second-level center in the transportation network and the next level of spoke is not involved.

3.4 Basic Assumptions

- Transport between start and end points need to pass two hub stations;
- Transportation mode from the point of supply to the hub is transported by road;
- Only one type of hub is built in the same province;
- When the hub transit distance is greater than 1.5 times the straight-through distance, the hub-spoke network will be abandoned and straight-through transport will be adopted;
- The transfer cost between the hubs is constant, and no transshipment costs are required in direct transport;
- Each station has adequate operation capacity;

- Piggyback transport vehicles have sufficient reserves to meet the needs of on-time operation.

3.5 Function Expression

The objective function of the optimization model:

$$Z = \min \sum_{i} \sum_{j} \sum_{k} \sum_{m} \mathrm{h}_{ij} X_{ij}^{km} C_{ij}^{km} + \sum_{i} \sum_{j} \sum_{k \neq j} \sum_{m \neq k \cup m \neq j} X_{ij}^{km} g_{\mathrm{ij}}^{km} \tag{1}$$

(1) *Total transportation cost*:

$$C_{ij}^{km} = C_{im}^{0} + C_{jk}^{0} + \alpha C_{km}^{1} \tag{2}$$

(2) In the transport route $i - j - k - m$*, if there is direct transport, there is no need for transfer. In this case, the node cost is not taken into account. In other cases, the transfer cost is calculated according to the number of nodes passing through the hub*:

$$g_{\mathrm{ij}}^{km} = \begin{cases} h_{ij} C_{ij}^{km} \\ \text{if } (k = i \cap m = j) \cup (k = m = j) \cup (k = m = i) \\ h_{ij} C_{ij}^{km} + g_k(\mathrm{h}_{ij}) \quad \text{if } k \neq i \cap (m{=}j \cup m{=}k) \\ h_{ij} C_{ij}^{km} + g_m(\mathrm{h}_{ij}) \quad \text{if } m \neq j \cap (k = i \cup m = k) \\ h_{ij} C_{ij}^{km} + g_m(\mathrm{h}_{ij}) + g_k(h_{ij}) \quad \text{else} \end{cases} \tag{3}$$

The constraint condition of the belt node cost location model based on axial and radial network:

(1) *The OD flow is required to be served by the hub station and pass through the hub station*:

$$\sum_{k} \sum_{m} X_{ij}^{km} = \partial_{ij} \quad \forall \mathrm{i,j} \tag{4}$$

(2) *Each pair of OD flow in the adoption of direct or axial and radial transport mode selection*:

$$\partial_{ij} = \begin{cases} 0, & 1.5 d_{ij} \geq d_{ij}^{km} \\ 1, & 1.5 d_{ij} \leq d_{ij}^{km} \end{cases} \tag{5}$$

(3) *Only when the node is selected as the hub station can the OD flow be regarded as being served by the node and carried into the calculation*:

$$\mathrm{Y}_k \geq X_{ij}^{km} \geq 0, \quad \forall\, i, \mathrm{j,k,}m \tag{6}$$

$$\mathrm{Y}_m \geq X_{ij}^{km} \geq 0, \quad \forall\, i, \mathrm{j,}k, m \tag{7}$$

(4) *The number of hub stations shall conform to the total number limit*:

$$\sum_k Y_k = P, \quad \forall k \tag{8}$$

(5) *The value limit of distribution between hub station and cargo flow line*:

$$Y_k = \{0, 1\}, \quad \forall k \tag{9}$$

$$X_{ij}^{km} \in [0, 1], \quad \forall\, i, j, k, m \tag{10}$$

(6) *When the transfer distance of the hub is more than 1.5 times the direct distance, the direct transport will be adopted*:

$$\beta = \frac{3d_{ij}}{2d_{ij}^{km}} - \left| \frac{3d_{ij}}{2d_{ij}^{km}} \right| \tag{11}$$

$$X_{ij}^{km} \leq \frac{3d_{ij}}{2d_{ij}^{km}} - \beta \tag{12}$$

4 Algorithm

In this paper, a heuristic algorithm is adopted to study the selection of points in multi-distributive axial-radial networks;

STEP 1: Determine the number of network nodes N, determine the number of hub nodes $P = \sqrt{N}$;

STEP 2: This paper proposes a new calculation index for node positioning of railway container transport network nodes planning hub nodes: $V = C/W$ Using the calculation results of each node in the network, the node corresponding to the smaller one is selected as the alternative hub anchor point. The small V value represents the node corresponding to the point with small transportation distance and large network flow from the network to other nodes, which has a good effect on the attracting ability of goods and the accessibility in the network;

STEP 3: Floyd algorithm is used to determine the distribution of each node and hub node in the network.

STEP 4: Calculated by using the most short-circuit path and transportation cost of network, pure axial radial network transport distance and cost, distance and transportation cost of hybrid shaft radial network sensitivity analysis, and respectively hub node set number in the debugging network, trunk line network transport discount, through transport and transit transport distance than indicators, such as to determine the optimal network setup parameters (Fig. 3).

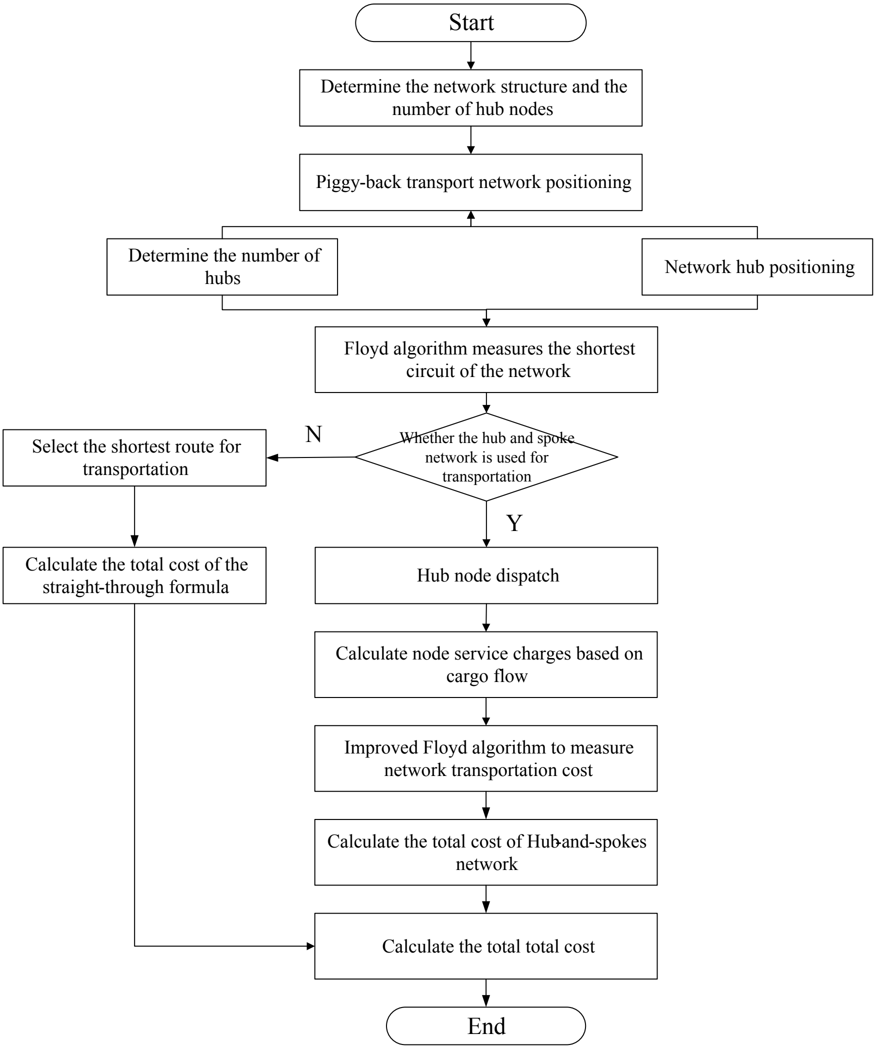

Start
Determine the network structure and the number of hub nodes
Piggy-back transport network positioning
Determine the number of hubs
Network hub positioning
Floyd algorithm measures the shortest circuit of the network
Whether the hub and spoke network is used for transportation
N
Y
Select the shortest route for transportation
Calculate the total cost of the straight-through formula
Hub node dispatch
Calculate node service charges based on cargo flow
Improved Floyd algorithm to measure network transportation cost
Calculate the total cost of Hub-and-spokes network
Calculate the total total cost
End

Fig. 3 Heuristic algorithm flow chart

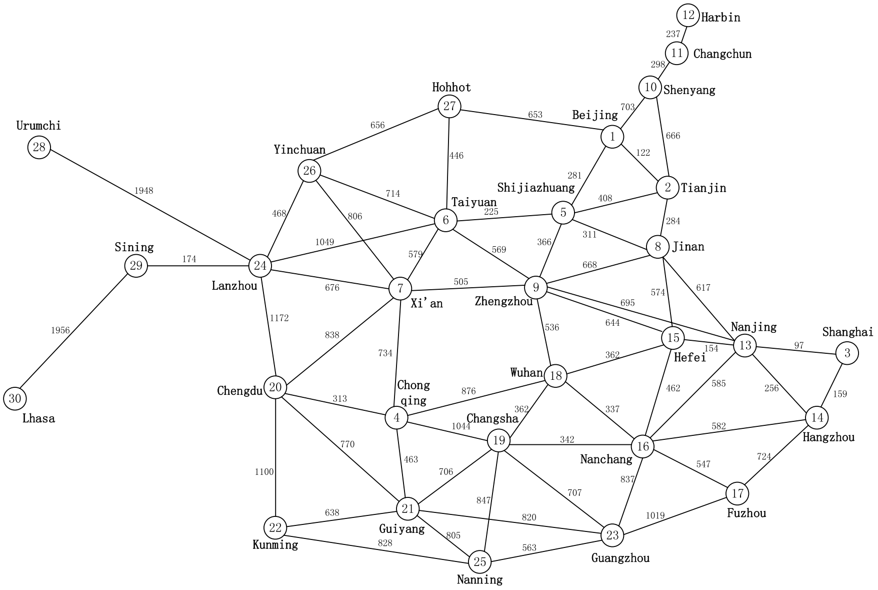


Fig. 4 Network diagram between alternative nodes

5 Computational Experiments

5.1 Case Background

In this paper, 33 first-level logistics centers that have been planned in China are selected as the alternative nodes to establish calculation examples (Fig. 4):

5.2 Modal Calculation

Based on the existing freight volume and the shortest transportation distance data, the calculation index of hub node positioning for railway container package-back network node planning is calculated as follows: $V = C/W$; The number of appropriate hubs is $\sqrt{N}$, and the number of selected nodes is 30. Therefore, the top 6 cities in index score are selected as the cities of alternative hubs.

5.3 Sensitivity Analysis

For the study of shaft radial network hub node number of changes to the total cost of the network effect, to maintain the existing transport network unit price and the

Table 1 The influence of the number of nodes on the total cost of the different networks

Node number	Hybrid axle-spoke network (Unit: ten thousand yuan)	Axial-radial network (Unit: ten thousand yuan)	Shortest circuit network (Unit: ten thousand yuan)
5	28,766,458	39,467,823	29,650,368
6	27,435,628	35,740,850	29,650,368
7	27,115,316	34,699,664	29,650,368
8	27,819,432	34,890,829	29,650,368
9	28,518,927	34,480,002	29,650,368
10	28,539,575	34,036,614	29,650,368
11	29,481,706	35,472,981	29,650,368
12	29,341,885	35,135,509	29,650,368

hub service fee is changeless, based on the existing six hub node number, for the hub node changes impact on the total cost of the network sensitivity analysis, in the process of increase and decrease the number of nodes, select hub has a better evaluation index node selection (Table 1).

5.4 Conclusion for Sensitivity Analysis

(1) According to the results of the change in hub station number, the total cost of network shows an "U" type change. When the hub node number is 7, the total cost reaches the minimum. As the number of hub nodes increase, the total cost to approach the cost of the short circuit network. This result shows that with the number of hub nodes increases, because of hub transit costs, forcing some flows that originally selected hub-and-spoke network convert to straight-through transport. The rising number of the hub nodes also make the scale of the main transport efficiency decreases.

(2) The current calculation example also shows that compared with the hybrid hub-and-spoke network, the total cost of the network has changed significantly during the change of the hub node compared to the pure hub-and-spoke network. This result may be due to all cargo flows are subject to hub transshipment in pure hub-and-spoke network. With the increase in the number of hub nodes, the cost of detours caused by transshipment is gradually diluted, and the scale benefits of the hub-and-spoke network are reflected. However, due to the impact of hub transshipment costs, The number of stations is gradually increasing, and the overall cost of the network has risen again due to the weakening of network scale benefits.

(3) From the analysis results, it can be seen that when the number of selected nodes is 7 nodes, the total network cost can be optimized, and the 7-node hub selection schemes are: **Chengdu, Guangzhou, Shenyang, Nanjing, Zhengzhou,**

Lanzhou, Changsha. After the number of node selection schemes gradually increased, the increase of hub nodes in the network originally formed a railway trunk line path between hub nodes that were originally close to each other, which weakened the advantages of the original railway in long-distance transportation.

6 Conclusion

Based on the idea of hub-and-spoke network, this paper designs a node planning model of container transport network and a heuristic algorithm to solve it. Based on the investigation data, the calculation example was analyzed, and the site selection scheme obtained from the calculation example was verified to obtain the current total network cost, and the change of the total network cost was studied according to the number of different hub nodes. It was found that the total network cost could be reduced to the minimum when 7 hub nodes are set.

Acknowledgements This paper was funded by the National Key R&D Program of China (grant number:2018YFB1201403) and Civil Aviation Administration of China Major projects (grant number: 201501).

References

1. J. Lin, A. Wang, C. Deng, D. Li, Development of railway container multimodal transport. Railway Freight **32**(01), 23–28 (2014)
2. B. Wu, *Research on the Technical Conditions for Carrying Out Piggy-Back Transportation in China Under the Existing Conditions* (Jiaotong university, Beijing, 2015)
3. E. Iain, *Ellis' British Railway Engineering Encyclopaedia* (Lulu, London, 2006), p. 293
4. N Yu, The development of American railway truck bumper. Foreign Railway Veh. (01), 25–28 (1985)
5. J. Peng, J. Si, F. Bao, Optimization of air network applied in the express based on hub-and-spoke network: a case study of SF express. *Computer, Informatics Cybernetics and Applications* (Springer, Netherlands, 2012), pp. 303–310
6. K. Yang, L. Yang, Z. Gao, Planning and optimization of intermodal hub-and-spoke network under mixed uncertainty. Transp. Res. Part E 95 (2016)
7. B. Wang, S. He, R. Song, Y. Shen, Traffic allocation optimization model and algorithm of fast freight network under integrated transportation system. J. Railway **31**(02), 12–16 (2009)
8. X. Lu, *Study on Economic Feasibility and Transportation Organization of China's Railway Carrying Out Piggy-Back Transportation* (Beijing Jiaotong University, 2015)
9. W. Najy, A. Diabat, Benders decomposition for multiple-allocation hub-and-spoke network design with economies of scale and node congestion. Transp. Res. Part B 133 (2020)
10. S. Liu, *Research on The Design of Highway Express Freight Axle and Spoke Network with Time Window* (Chang 'an University, 2011)
11. F. Alkaabneh, A. Diabat, S. Elhedhli, A Lagrangian heuristic and GRASP for the hub-and-spoke network system with economies-of-scale and congestion. Transp. Res. Part C, 102 (2019)
12. S.T. Choong, M.H. Cole, E. Kutanoglu, Empty container management for intermodal transportation networks. Transp. Res. Part E **38**(6), 423–438 (2002)

13. S. Yan, C.H. Lai, H.C. Chen, Ashort-term fright scheduling model for international express package delivery. J. Air Transp. Manage **11**(6), 368—374 (2005)
14. L.-L. Ni, F. Shi, Optimization method of hub location and distribution for multi-distribution express hub and hub network. Syst. Eng. Theor. Pract. **32**(02), 441–448 (2012)
15. *Study on Site Selection and Route Optimization of Hub and Spoke Logistics Network Design* (Huazhong University of Science and Technology, 2007)
16. J. Fan, Y. Lu, Y. Xie, X. Xu, Construction and empirical study of regional logistics channel based on axial amplitude network. Logistics Technol. **33**(13), 205–207 (2014)

Simulation and Optimization of Automated Warehouse Based on Flexsim

Xiang Li, Li Wang, and Xiaoning Zhu

Abstract In this paper, in order to reduce the logistics costs and commodity storage costs in the actual operation of e-commerce enterprises, we introduce an automated three-dimensional warehouse in the e-commerce logistics storage process, and use Flexsim software to model the ASRS system, discrete event simulation, system optimization. Afterwards, the bottleneck analysis and optimization of the logistics sorting process and processing outflow process are carried out to achieve the effect of making e-commerce logistics warehousing operations operate efficiently. Through this article, e-commerce enterprises can learn from the logistics warehouse operation process and process optimization steps.

Keywords Simulation · Automated warehouse · Flexsim

1 Introduction

With the development of economy and society, the Internet and big data are used more and more widely, and e-commerce has emerged with the advancement of technology and network. The new social form and development model have put forward new requirements for e-commerce enterprises to achieve perfect and fast connection of the entire supply chain. The logistics system process is indispensable for scale and systematization, and warehousing is a key link in the logistics process. Logistics cost compression plays a vital role. Effective warehousing management can affect the intensification and effectiveness of the entire supply chain to a certain extent, fully reduce the impact of the bullwhip effect, and promote the sustainable development of e-commerce logistics.

X. Li (✉) · L. Wang · X. Zhu
School of Transportation Engineering, Beijing Jiaotong University, Beijing, China
e-mail: 19125738@bjtu.edu.cn

L. Wang
e-mail: liwang@bjtu.edu.cn

X. Zhu
e-mail: xnzhu@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_18

At present, domestic e-commerce companies still have the problems of extensive warehouse management and inefficient warehouse management, asymmetry and inconsistency of information between upstream and downstream suppliers, and high inventory costs, disordered storage procedures, and explosion of network product storage. In order to effectively solve the above problems, we must first solve the management efficiency and operation efficiency of the logistics storage area, and improve the flexibility of the entire storage business. This article is organized as follows. The second chapter mainly introduces related research on logistics warehousing optimization. In the third part, an automated 3D warehouse model based on flexsim is established, the inbound sorting and outbound processing flow is simulated, the simulation results are analyzed, and the model is optimized. Carry out the second simulation in Sect. 4. In Sect. 5, we pointed out the defects in the model and looked forward to the future research direction.

2 Literature Review

At present, there are two main directions for the research and development of logistics system modeling and simulation in China. One case is that when the logistics system of the research is not very complicated, or the complexity of the system is reduced, the mathematical methods such as linear algebra can be used. Calculus, operations research, computational mathematics, etc. to solve the problem.

Tian Congress, Zhang Pan (2004) uses hybrid genetic algorithm to optimize fixed shelf selection [1]. Later, Shu Hui (2018) optimized and supplemented the Petri net-based modeling method in his own article [2]. Meng Kui (2014) used Flexsim software to create a more complex warehouse operation model, introduced multiple network nodes, planned a fixed shipping order, and analyzed the feasibility of the solution by selecting the results of different sequential schemes. At the same time, Meng Kui introduced a new one. The idea is to use the logistics network between the working equipment to represent the information of different places, and optimize the shortest path with time as the weight [3].

Niu Jin (2015) used the food supply as a case to study the logistics operation under the emergency situation. This kind of logistics emergency situation, which is one of the common situations in logistics operations, is usually obvious in e-commerce enterprises. The incident has been prepared accordingly, and the decision can be made quickly when the incident occurs. The three elements in the emergency logistics system are emergency reserve, emergency processing and emergency dispatch, and the emergency system is simulated by Flexsim. Solve the weak links in the emergency system [4].Zheng Zhenyu, Yuan Hongbing and Xu Zhenglin (2019) of Nanjing Liuwei Logistics Equipment Co., Ltd., based on SimPy modeling and simulation of AS/RS, the modeling process is mainly based on process, combined with Python language to build logical connections between processes. This method belongs to a novel method in the modeling and simulation of logistics system [5]. In the current NilsBoysen, Renede Koster, Felix Weidinger (2018) surveyed the storage area in the e-commerce era, he roughly divided the storage area into the following, the traditional single-piece storage area, high-rise mixed shelf storage area, quantitative Partitioned storage area, automated three-dimensional storage area [6]. Shu Hui (2018) optimized and improved the modeling method based on Petri net in his article [7].

Zhao Lingxing, Zhong Liangwei (2018) focused on the model optimization of the automated three-dimensional storage area under the retail e-commerce platform, using improved genetic algorithms, setting new individual chromosome codes, creating fitness functions and genetic operators for simulation [8] Zheng Zhenyu, Yuan Hongbing and Xu Zhenglin (2019) of Nanjing Six-Dimensional Logistics Equipment Co., Ltd. (SIM) modeled and simulated AS/RS based on SimPy. The modeling process is mainly based on the process, combined with the Python language to build the logical connection between the processes, This method belongs to a relatively novel method in logistics system modeling and simulation [9].

3 Optimization of AS/RS Simulation Based on Flexsim

In the overall modeling process, simulate the effects of the initial storage area according to the following simulation components, as shown in the Table 1:

3.1 Inbound Model

Then create a logical connection, a generator and a staging area, a conveyor belt, create an A connection between the shelves in turn, create an A connection between the robot and the task distributor, create an S connection between the temporary

Table 1 Simulation effect of each component

Simulation component	Simulation effect
Source	Three corresponding generator simulations correspond to upper-level supplier delivery
Queue	Simulate the actual cargo temporary storage area
Conveyor	Simulated AS/RS cargo transport instead of manual operation
Rack	Unity attic shelf using high-rise shelves to simulate AS/RS
Robot	Automatic handling process in simulated AS/RS
ASRSvehicle	Simulate the loading and unloading process of loading and unloading goods
Transportor	Analog automatic guided trolley
Combiner	Packing table for simulating shipment processing
Sorting conveyor	Simulate shipment sorting

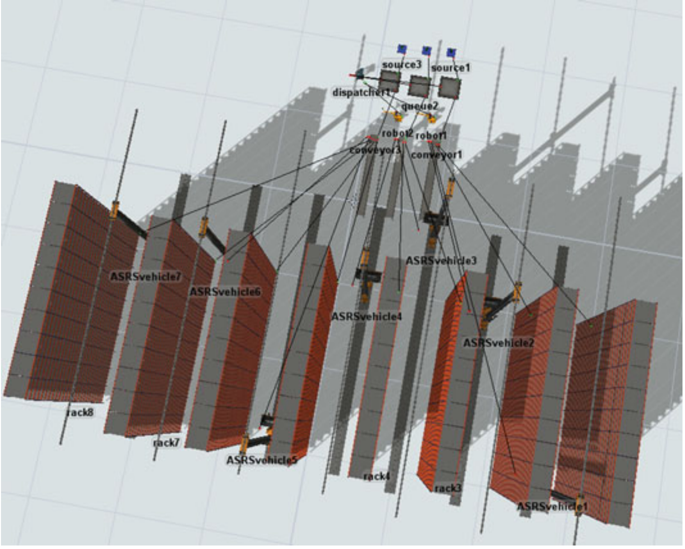


Fig. 1 The operation modes of the sending and arrival containers

storage area and the task distributor, and a task distributor The principle of allocation is to select the idle unoccupied port. Subsequently, the first and second shelves are ready to accept the red entity, the third to fifth shelves receive the yellow entity, the sixth to eighth shelves accept the blue entity, and the lane between each two shelves is set to one. The stacking machine is used for handling operations. The initial warehousing model is as follows (Fig. 1).

3.2 Processing Outbound Model

Drag and drop high-rise shelves, temporary storage areas, forklifts (using forklifts to simulate automatic guided cars), generators, synthesizers, sorting conveyors, etc. in Flexsim, where three high-rise shelves are filled with three entities For example, the simulation simulation outbound processing operation, the maximum storage volume per shelf is 260, so the generator is connected to each shelf A, and the arrival parameters of the generator are set to three types of entities, the limited number is 260 And select the output port according to the type, and set its own trigger to set the corresponding color to red, yellow, blue.

The task distributor is used to distribute the forklift, the distribution principle selects the cycle mode, the A connection is made between the shelf and the synthesizer, and the generator connected to the synthesizer A mainly produces five types of pallets, symbolizing five different customer orders, in the global table. Set the different requirements of the corresponding customer orders for the three entities,

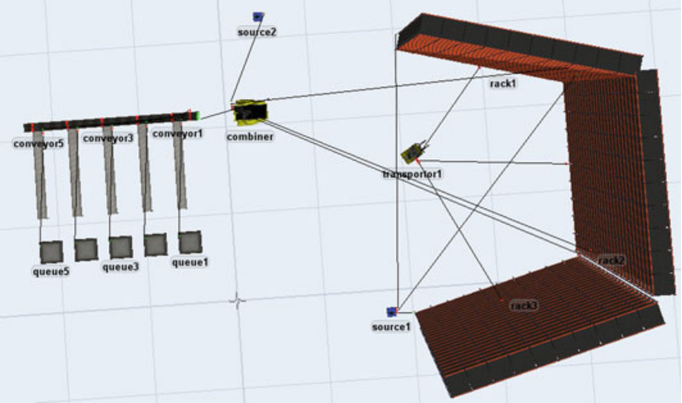


Fig. 2 Schematic diagram of processing outbound model

and bring the global table into the synthesizer. The synthesizing principle of the synthesizer is synthesized, according to the table. According to the type, the specific model diagram is shown in Fig. 2.

4 Model Analysis and Optimization

4.1 Inbound Results Analysis Optimization

Before running the inbound model, the running time is set to 20,000 s. After the running is finished, the statistics bar is opened, and the running result is transmitted in excel form. The data is as follows (Table 2).

Analyze the working efficiency of the robot and the roadway stacker. The set time is 20000 s, and the corresponding no-load and full-load operation time rate are calculated, which is roughly as shown in the Fig. 3 (Table 3).

It can be seen from the above data that the operating efficiency of the No. 1 robot and the No. 2 robot are not high, and there is a large amount of idle time. At the same time, the ASRSvechical of No. 2, 4, 6, and 7 have low working efficiency and are basically idle.

So improve the model, set up a robot, remove the ASRSvechical and get the data output as follows (Table 4).

More fully demonstrate the working effect of each device through the pie chart (Fig. 4).

After the Dashboards test, it can be determined that the idle rate of the corresponding transportation equipment is relatively balanced, the optimization result is considerable, and the secondary simulation is successful.

Table 2 Inbound job simulation data output

Flexsim Summary Report							
Time:	20,000						
Object	Class	stats_content	stats_contentmin	stats_contentmax	stats_contentavg	stats_input	stats_output
source1	Source	0	0	0	1	0	250
source2	Source	0	0	0	1	0	200
source3	Source	0	0	0	1	0	401
robot1	Robot	1	0	1	0.040008	161	160
robot2	Robot	0	0	1	0.157529	630	630
rack2	Rack	133	1	133	64.38601	133	0
rack3	Rack	99	1	99	44.61271	99	0
rack4	Rack	100	1	100	53.2381	100	0
rack5	Rack	78	1	78	36.89861	78	0
rack6	Rack	68	1	68	28.25892	68	0
rack7	Rack	88	1	88	46.22017	88	0
rack1	Rack	115	1	115	58.04571	115	0
rack8	Rack	90	1	90	45.98136	90	0
ASRSvehicle5	ASRSvehicle	1	0	1	0.495973	325	324
ASRSvehicle7	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle6	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle4	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle3	ASRSvehicle	0	0	1	0.238281	199	199
ASRSvehicle2	ASRSvehicle	0	0	0	0	0	0

(continued)

Table 2 (continued)

Flexsim Summary Report							
Time:	20,000						
ASRSvehicle1	ASRSvehicle	1	0	1	0.442861	249	248
queue1	Queue	1	0	1	0.062914	250	249
queue2	Queue	0	0	1	0.050802	200	200
queue3	Queue	59	0	60	25.17327	401	342

Flexsim Summary Report						
Time:	20,000					
Object	Class	stats_staytimemin	stats_staytimemax	stats_staytimeavg	state_current	state_since
source1	Source	0	0	0	5	19,997.19
source2	Source	0	0	0	5	19,991.11
source3	Source	0	0	0	5	19,977.5
robot1	Robot	5	5	5	17	19,996.11
robot2	Robot	5	5	5	16	19,997.19
rack2	Rack	0	0	0	1	0
rack3	Rack	0	0	0	1	0
rack4	Rack	0	0	0	1	0
rack5	Rack	0	0	0	1	0
rack6	Rack	0	0	0	1	0
rack7	Rack	0	0	0	1	0
rack1	Rack	0	0	0	1	0

(continued)

Table 2 (continued)

Flexsim Summary Report						
Time:	20,000					
rack8	Rack	0	0	0	1	0
ASRSvehicle5	ASRSvehicle	14.45579	50.81482	30.59412	17	19,985.97
ASRSvehicle7	ASRSvehicle	0	0	0	1	0
ASRSvehicle6	ASRSvehicle	0	0	0	1	0
ASRSvehicle4	ASRSvehicle	0	0	0	1	0
ASRSvehicle3	ASRSvehicle	15.37427	33.83937	23.90085	1	19,960.8
ASRSvehicle2	ASRSvehicle	0	0	0	1	0
ASRSvehicle1	ASRSvehicle	26.9575	45.07735	35.66003	17	19,969.43
queue1	Queue	5	10.62962	5.052625	10	19,997.19
queue2	Queue	5	10.47272	5.079248	6	19,996.11
queue3	Queue	5	2951.771	1220.211	8	19,977.5

Flexsim Summary Report							
Time:	20,000						
Object	Class	idle	processing	blocked	generating	empty	utilize
source1	Source	0	0	0	19,997.19	0	
source2	Source	0	0	0	19,991.11	0	
source3	Source	0	0	0	19,977.5	0	
robot1	Robot	18,391.11	0	0	0	0	0
robot2	Robot	13,697.19	0	0	0	0	0

(continued)

Table 2 (continued)

Flexsim Summary Report							
Time:	20,000						
rack2	Rack	0	0	0	0	0	
rack3	Rack	0	0	0	0	0	
rack4	Rack	0	0	0	0	0	
rack5	Rack	0	0	0	0	0	
rack6	Rack	0	0	0	0	0	
rack7	Rack	0	0	0	0	0	
rack1	Rack	0	0	0	0	0	
rack8	Rack	0	0	0	0	0	
ASRSvehicle5	ASRSvehicle	180.3188	0	0	0	0	0
ASRSvehicle7	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle6	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle4	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle3	ASRSvehicle	11,860.17	0	0	0	0	0
ASRSvehicle2	ASRSvehicle	0	0	0	0	0	0
ASRSvehicle1	ASRSvehicle	3420.659	0	0	0	0	0
queue1	Queue	0	0	0	0	18,739.08	
queue2	Queue	0	0	0	0	18,980.26	
queue3	Queue	0	0	0	0	3180.746	

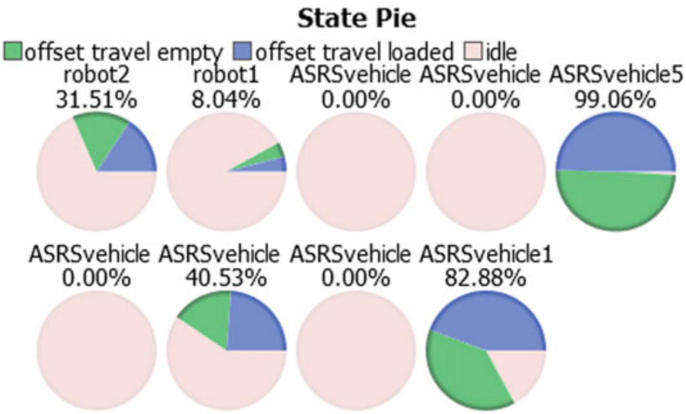


Fig. 3 Equipment idle and utilization

Table 3 Equipment idle and utilization

Flexsim Summary Report								
Time:	20,000							
Object	Class	idle	blocked	travel empty	travel loaded	offset travel empty	offset travel loaded	utilize
robot1	Robot	18,391.11	0	0	0	805	800	0
robot2	Robot	13,697.19	0	0	0	3150	3150	0
ASRSvehicle5	ASRSvehicle	180.3188	0	0	0	9893.155	9912.494	0
ASRSvehicle7	ASRSvehicle	0	0	0	0	0	0	0
ASRSvehicle6	ASRSvehicle	0	0	0	0	0	0	0
ASRSvehicle4	ASRSvehicle	0	0	0	0	0	0	0
ASRSvehicle3	ASRSvehicle	11,860.17	0	0	0	3344.353	4756.27	0
ASRSvehicle2	ASRSvehicle	0	0	0	0	0	0	0
ASRSvehicle1	ASRSvehicle	3420.659	0	0	0	7705.084	8843.687	0

Table 4 Optimized equipment operation

Flexsim Summary Report								
Time:	20,000							
Object	Class	idle	blocked	travel empty	travel loaded	offset travel empty	offset travel loaded	utilize
robot1	Robot	12,031.13	0	0	0	3985	3980	0
ASRSvehicle5	ASRSvehicle	108.3081	0	0	0	9922.866	9949.497	0
ASRSvehicle3	ASRSvehicle	11,952.49	0	0	0	3306.903	4699.918	0
ASRSvehicle1	ASRSvehicle	3776.545	0	0	0	7531.97	8672.483	0

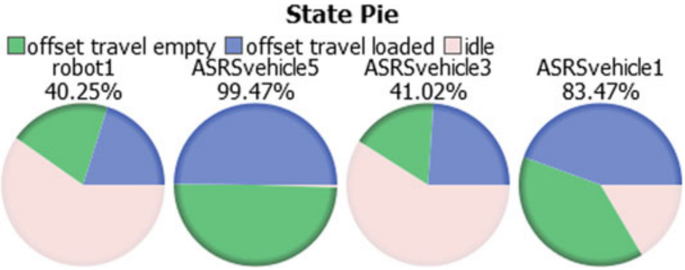


Fig. 4 Equipment idle and utilization after optimization

4.2 Analysis and Optimization of Processing Outbound Model

The data obtained by running the original machining outbound model is shown below (Table 5):

Use a pie chart to indicate the efficiency of the forklift, the specific shape is as follows (Fig. 5):

As can be seen from the figure, the forklift is too busy to work, and the work arrangement is unreasonable,so we improve the model and add two transportors, subsequent simulations are performed to obtain the data sheet shown below (Table 6):

Converted into a better visible pie chart as follow (Fig. 6):

The analysis shows that the usage of the forklift is relatively balanced, the working time is not much different, and there is no waste of resources. In the process of operation, since the setting amount of the global table is five, the order is stopped after the five finished entities are synthesized. Therefore, the problem of this model is that the running time is short, which is easy to cause the problem of insufficient persuasiveness of the running result. Therefore, optimization is performed on the basis of this, and the order quantity of the global table is changed. In the actual situation, the phenomenon is often a certain For a sudden demand of a kind or a certain type of production, only complete global table information is needed to ensure that the model can ensure the normal operation of processing and delivery in an emergency situation. And we can see that the work efficiency of the three forklift trucks is relatively concentrated, and the work distribution is relatively average. The secondary simulation effect is better.

In summary, comparing the device usage rates of the two models before and after optimization, the following table is obtained (Tables 7 and 8):

Table 5 Equipment idle and utilization after optimization

Flexsim Summary Report											
Time:	2183.464										
Object	Class	stats_content	stats_input	stats_output	stats_staytimemin	stats_staytimemax	stats_staytimeavg	idle	processing	blocked	utilize
rack1	Rack	87	100	13	7.553331	1509.224	545.8866	1509.224	0	0	0
rack2	Rack	80	100	20	76.87309	1740.223	883.0307	1740.223	0	0	0
rack3	Rack	76	100	24	226.1133	2124.188	1262.227	2124.187	0	0	0
source1	Source	0	0	300	0	0	0	0	0	0	0
transportor1	Transporter	0	57	57	9.390796	20.277	15.40112	40	0	0	0
combiner	Combiner	0	62	62	0	693.1705	34.74943	0	50	0	0
queue5	Queue	1	1	0	0	0	0	0	0	0	0
queue4	Queue	1	1	0	0	0	0	0	0	0	0
queue3	Queue	1	1	0	0	0	0	0	0	0	0
queue2	Q ueue	1	1	0	0	0	0	0	0	0	0
queue1	Queue	1	1	0	0	0	0	0	0	0	0
source2	Source	0	0	5	0	443.9538	292.2588	0	0	1461.294	0

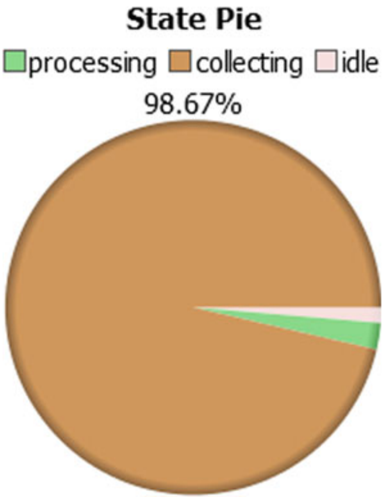


Fig. 5 Forklift operation efficiency and idle rate

5 Conclusion

The main conclusion drawn in this article is that the automated three-dimensional warehouse has a good effect in logistics warehousing operations. It has certain advantages in handling large batches and multi-batch tasks and orders. At the same time, the automated three-dimensional warehouse also has certain space and process optimization in the process of operation and resource allocation. Through logistics system simulation, it can effectively solve low efficiency, The problem of high inventory and high cost. Through the use of Flexsim software to simulate logistics and warehousing operations, the 3D effect of actual warehousing operations is further demonstrated, and bottleneck points and perfectable points in warehousing operations are searched through data. However, there are still some defects in the research process. First of all, the order setting only involves Poisson orders and global phenotype orders, and does not consider the actual order explosion. The model focuses on optimizing the operation process and resource allocation level. The overall optimization of the chain system has certain limitations, and future research will start from the above two aspects.

Table 6 Optimized data output

Flexsim Summary Report									
Time:	837.2989								
Object	Class	stats_content	stats_contentmin	stats_contentmax	stats_contentavg	stats_input	stats_output	stats_staytimemin	stats_staytimemax
rack1	Rack	87	1	100	91.71896	100	13	9.818585	575.9686
rack2	Rack	80	1	100	89.99211	100	20	40.33258	663.2368
rack3	Rack	76	1	100	90.326	100	24	76.10732	778.0219
source1	Source	0	0	99	100	0	300	0	0
dispatcher	Dispatcher	0	0	0	0	0	0	0	0
transportor1	Transporter	0	0	1	0.369652	19	19	9.590977	19.577
transportor2	Transporter	0	0	1	0.381853	19	19	9.607718	20.277
transportor3	Transporter	0	0	1	0.375004	19	19	9.569872	18.577
combiner	Combiner	0	0	2	1	62	62	0	252.4144
queue5	Queue	1	1	1	0	1	0	0	0
queue4	Queue	1	1	1	0	1	0	0	0
queue3	Queue	1	1	1	0	1	0	0	0
queue2	Queue	1	1	1	0	1	0	0	0
queue1	Queue	1	1	1	0	1	0	0	0
source2	Source	0	0	0	1	0	5	0	170.8887

Flexsim Summary Report								
Time:	837.2989							
Object	Class	stats_staytimeavg	state_current	state_since	idle	blocked	processing	utilize
rack1	Rack	209.0749	1	575.9685	575.9685	0	0	0

(continued)

Table 6 (continued)

Flexsim Summary Report								
Time:	837.2989							
rack2	Rack	331.3568	1	663.2367	663.2367	0	0	0
rack3	Rack	464.4141	1	778.0219	778.0219	0	0	0
source1	Source	0	5	0	0	0	0	0
dispatcher	Dispatcher	0	1	0	0	0	0	0
transportor1	Transporter	15.12403	1	777.3712	86.48262	0	0	0
transportor2	Transporter	16.04383	1	798.2989	68.98417	0	0	0
transportor3	Transporter	15.341	1	777.2698	79.25748	0	0	0
combiner	Combiner	13.03708	1	808.2989	0	0	50	0
queue5	Queue	0	8	119.2165	0	0	0	0
queue4	Queue	0	8	248.678	0	0	0	0
queue3	Queue	0	8	405.9958	0	0	0	0
queue2	Queue	0	8	580.8845	0	0	0	0
queue1	Queue	0	8	837.2989	0	0	0	0
source2	Source	111.1769	5	555.8845	0	555.8845	0	0

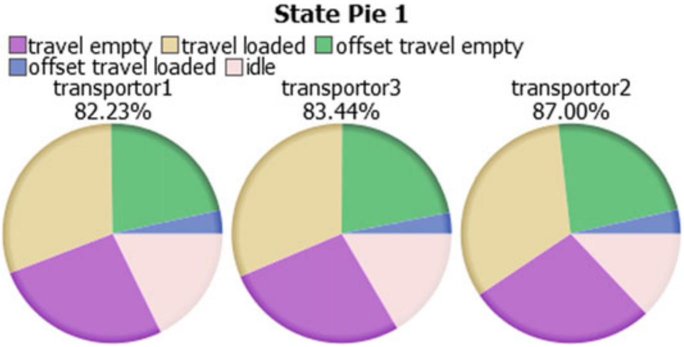


Fig. 6 The pie of optimized data output

Table 7 Comparison of equipment utilization about warehousing operation optimization

Evaluation	No-load time rate before optimization (%)	No-load time rate after optimization (%)	Idle time rate before optimization (%)
Equipment			
Robot			70.2
Stacker1	32.47	28.47	47.86
Stacker2	16.04	15.82	71.22
Stacker3	29.06	23.13	56.19

Table 8 Processing and processing equipment utilization comparison

Evaluation	Idle time rate before optimization(%)	Idle time rate after optimization(%)
Equipment		
Transportor1	1.79	9.88
Transportor2		9.72
Transportor3		10.18

Acknowledgements This paper was funded by Civil Aviation Administration of China Major projects (grant number: 201501) and the National Key R&D Program of China (grant number: 2018YFB1201403) .

References

1. Tian, G., Zhang, P., Research on the optimization problem of fixed shelf selection based on hybrid genetic algorithm. Mech. Eng. Rep. (40), 141–144 (2004)
2. Shu, H., Modeling and optimization of automated three-dimensional storage area based on Petri net. Manuf. Autom. **40**(11), 148–149+156 (2018)

3. Meng, K., Research on simulation and optimization of logistics operation system based on Flexsim. Logistics Technol. **33**(09), 422–423+438 (2014)
4. Niu, J., *Simulation and Optimization of Food Emergency Logistics System Based on Flexsim* (Henan University of Technology, 2015)
5. Zheng, Z., Yuan, H., Xu, Z., Simulation and simulation of automated stereo storage area based on SimPy. Manuf. Autom. **41**(03), 102–105+108 (2019)
6. Boysen, N., Koster, R., Weidinger, F., Warehousing in the e-commerce era: a survey. Eur. J. Oper. Res. Available online 2018
7. Shu, H., Modeling and optimization of automated three-dimensional warehouse based on Petri net. Manuf. Autom. **40**(11), 148–149+156 (2018)
8. Z. Lingxing, Z. Liangwei, Optimization simulation of intelligent warehousing system based on Flexsim. Softw. Guide **17**(04), 161–163 (2018)
9. Zheng, Z., Yuan, H., Xu, Z., Modeling and simulation of automated three-dimensional warehouse based on SimPy. Manuf. Autom. **41**(03), 102–105+108 (2019)

Research on Path Planning of Logistics Storage Robot Based on Fuzzy Improved Artificial Potential Field Method

Guanyi Liu, Yanping Du, Xinyue Li, and Shuihai Dou

Abstract In the complex and dynamic environment of logistics storage, obstacles have randomness and uncertainty. The traditional artificial potential field method has some problems in the process of path planning, such as inaccessibility and poor real-time performance, which can not meet the working performance requirements of logistics storage robot. In order to solve the problem that the target of traditional artificial potential field method is not reachable, the original repulsion potential field function is improved by introducing a distance adjustment factor to help the logistics storage robot reach the target point smoothly; then the new repulsion function is obtained by introducing the relative speed and acceleration between the robot and the obstacle, and the coefficient of repulsion function is adjusted in real time by combining the fuzzy logic control algorithm. Finally, the simulation experiment is carried out by MATLAB. The experimental results show that the artificial potential field method is feasible and effective in path planning.

Keywords Artificial potential field method · Fuzzy logic control · Logistics storage robot · Path planning

G. Liu · Y. Du · S. Dou (✉)
School of Mechanical and Electrical, Beijing Institute of Graphic Communication, Beijing, China
e-mail: doushuihai@126.com

G. Liu
e-mail: 15501092112@163.com

Y. Du
e-mail: duyanping@bigc.edu.cn

X. Li
School of Architecture and Transportation, Guilin University of Electronic Technology, Guilin, China
e-mail: 3514823386@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_19

1 Introduction

With the rapid development of the Internet and e-commerce, there is an increasing demand for the improvement of logistics efficiency. Intelligent logistics warehousing robot has been widely used in the book, manufacturing, footwear, e-commerce and other industries, and new logistics warehousing robot design schemes are also emerging. In terms of factory production logistics, with the concept of "made in China 2025" proposed in the national high-tech strategy, the manufacturing industry has taken the step of transition to intelligence and is in the upsurge of building 'intelligent production workshops" [1]. Thus, it can be seen that the demand for flexible indoor logistics in the industrial scene has gradually matured, which makes it important for logistics warehousing robot to improve the accuracy of route planning and intelligent obstacle avoidance function. The complexity of the working environment of logistics warehouse robot determines that the performance of path planning is an important index to measure the performance of logistics warehouse robot.

Path planning is one of the main functions of logistics warehousing robot. It means that in an environment with obstacles, the robot searches for an optimal or suboptimal non-collision path from the starting point to the target point according to certain evaluation criteria [2]. The quality of path planning is related to whether the robot can successfully complete the preset tasks safely and without collision. By improving the performance of path planning, the operation efficiency of mobile robot in complex environment can be effectively improved. Since the end of the nineteenth century and the twentieth century, many scholars at home and abroad have studied the path planning of indoor mobile robots and proposed many effective intelligent planning algorithms, such as bp artificial neural network [3] and particle swarm optimization algorithm [4]. The artificial potential field method was first proposed by Khatib [5], which is to construct an abstract potential field artificially in the system space. The mobile robot generates motion in the working space under the influence of the potential field force to complete the path planning. The principle of artificial potential field method is simple and easy to implement. It has good characteristics in local obstacle avoidance and path smoothing of mobile robot [6]. However, this algorithm has some inherent defects. In the in-depth study of artificial potential field method, scholars at home and abroad have proposed roughly four kinds of solutions. The first is to improve the gravity function and repulsion function to solve the problem of collision and target unreachable when the initial gravity is too large. Secondly, under the premise of using the original potential field model, the local minimum of target unreachable or escape can be solved. Thirdly, it integrates with other algorithms to learn from each other. The fourth is hierarchical control. Fazli and Kleeman [7] solved the problem of artificial potential field method by using obstacle avoidance algorithm along the wall. Chen et al. [8] analyzed the connectivity of obstacles and set up temporary target points to help the robot get out of the minimum state, effectively reducing the search space and improving the search efficiency. Ying [9] makes the robot out of the oscillating state by randomly generating temporary target points when it is at a local minimum. However, if the temporary target points are randomly placed inside the obstacles, the robot cannot escape. Liu et al. [10] improved the repulsion function by decomposing the repulsion into two directions and redefining it, thus solving the local minimum problem. Huo et al. [11] proposed a method of randomly changing the target point. By randomly changing the target point to change the robot's force, the resultant force is not zero when the robot is trapped in the local pole, so that the robot can escape from the local minimum. By setting up virtual intermediate target points and obstacles, Zhang and Zeng [12] changed the robot's force so that the robot could reach the target point smoothly. Li et al. [13] tested different obstacle repulsion coefficient, gravity coefficient and obstacle influence distance, and selected the optimal parameters for path planning through comparative analysis. However, when the robot's walking environment changed, it needed to adjust the parameters according to the new environment, which failed to fundamentally solve the problem. Meng et al. [14] proposed a fuzzy artificial potential field method based on dynamic of soccer robot path planning method, through fuzzy decision to adjust the repulsive force parameters such as the robot in each moment force size and direction of change, avoid robots into local minimum point, but the face of complex obstacle groups around cannot effectively out of trap area; Liu [15] introduced euclidian distance into the repulsion force and proposed an obstacle connection method based on situational behavior, which solved the problem of target unreachable and local minimum on the whole. However, in the u-shaped obstacles, the robot would wander and fail to reach the target point. In view of the fact that robots are prone to violent shocks and local optimality in complex and changeable environments, Han and Sun [16] proposed a fuzzy artificial potential field method for fuzzy decision processing of potential field parameters, but it failed to completely overcome the problem of violent shocks.

In order to solve the problem of traditional artificial potential field method, a fuzzy improved artificial potential field method is proposed in this paper. Aiming at the problem of target unreachable, the original repulsion potential field function is improved by introducing a distance regulator. In order to enable the robot to cope with the complex and dynamic logistics and storage environment, the relative speed and acceleration of the robot and the obstacle are introduced, and the three coefficients of the improved repulsion function are adjusted in real time by combining the fuzzy logic control algorithm, and the comprehensive decision is made to achieve efficient obstacle avoidance. Finally, through MATLAB simulation experiment to verify.

2 Traditional Artificial Potential Field Method

The artificial potential field method was first proposed by Khatib in 1986 to solve the obstacle avoidance problem of arm robots in motion. Later, researchers found that the algorithm had good performance in the field of path planning of mobile robots and was widely used in this field [17]. The basic idea of the artificial potential field method is to transform the whole environment into a large virtual artificial potential field model, abstract the mobile robot into a point charge moving in the

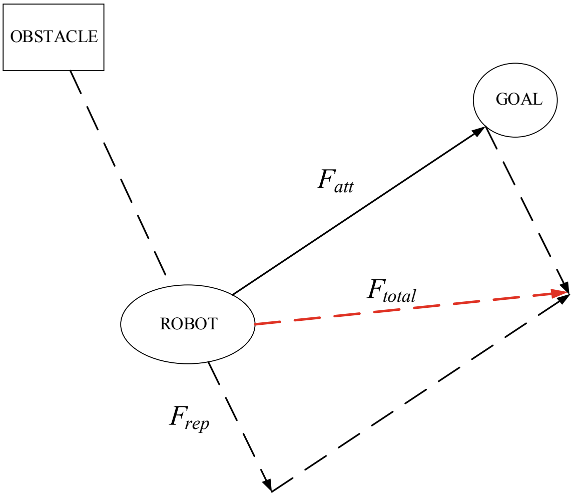


Fig. 1 Stress analysis of traditional artificial potential field method

virtual artificial potential field, and generate the attractive potential field at the target point and the repulsive potential field at the obstacle [18, 19]. When the robot moves in the potential energy field, it is affected by two forces: one is the attraction of the gravitational potential field of the target point to the robot, which makes the robot move towards the target point; the other is the repulsion force of the repelling potential field of the obstacles in the environment, which makes the robot move away from the obstacles. The force analysis diagram of the robot is shown in Fig. 1.

The robot is simplified to a model, and its motion space is two-dimensional, assuming that the position of the robot is $X_R = (x, y)$, Then the gravitational field function X_R of the target position to point $U_{\text{att}}(X)$ is:

$$U_{\text{att}}(X)=\frac{\eta}{2}d^2(X_R - X_G) \tag{1}$$

where $d(X_R - X_G)$ represents the Euclidean distance between the robot and the target, X_G is the location of the target point, η is an adjustable constant representing the coefficient of the gravitational field function.

The corresponding gravitational force produced by a function of the gravitational potential field is $F_{\text{att}}(X)$:

$$F_{\text{att}}(X) = -\nabla U_{\text{att}}(X) = -\eta d(X_R - X_G)e_{\text{RG}} \tag{2}$$

where e_{RG} represents the unit vector pointing from the current position of the logistics warehousing robot to the target position. According to Eqs. (1) and (2), the further the target position is from the current position of the logistics warehousing robot, the greater the gravitational potential will be, and the greater the gravity of the target position on the logistics warehousing robot will be. On the contrary, the closer the logistics warehousing robot is to the target, the lower the gravitational potential energy will be, and the lower the corresponding gravitational force will be. When the gravitational potential energy is zero, it indicates that the mobile robot is at the target position.

The repulsion field function between the logistics warehousing robot and the obstacle is defined as:

$$U_{\text{rep}}(X) = \begin{cases} \frac{\xi}{2}\left[\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\text{lim}}}\right]^2 & d(X_R - X_O) \le d_{\text{lim}} \\ 0 & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{3}$$

where $d(X_R - X_O)$ represents the Euclidean distance between the robot and the obstacle, d_{lim} is a positive constant representing the distance of influence of an obstacle, and ξ is an adjustable constant representing the proportional coefficient of the repulsive potential field function. X_O is the position of the nearest obstacle to the robot.

The negative gradient of the repulsion potential field function is the repulsion force suffered by the robot:

$$F_{\text{rep}}(X) = \begin{cases} \xi\left[\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\text{lim}}}\right] \cdot \frac{e_{\text{OR}}}{d^2(X_R - X_O)} & d(X_R - X_O) \le d_{\text{lim}} \\ 0 & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{4}$$

where e_{OR} represents the unit vector pointing from the obstacle to the current position of the logistics warehousing robot.

The resultant force received by the robot in the logistics storage environment $F_{\text{total}}(X)$ can be expressed as:

$$F_{\text{total}}(X) = F_{\text{att}}(X) + F_{\text{rep}}(X) \tag{5}$$

Artificial potential field method is a very effective static path planning method, but it has some defects in the dynamic environment with multiple obstacles. Through the analysis of the traditional artificial potential field method, it can be seen that due to the diversity of the movement mode, the shape of the obstacles, the target points and the relative positions of the obstacles of the logistics warehousing robot, the problem of unreachable targets in the process of path planning mainly has the following centralized typical cases:

(1) When the target exists around obstacles, as logistics warehousing robot near target, logistics warehouse robot by gravity decreases under obstacles repulsion is gradually increasing, when there is obstruction of logistics warehouse robot repulsion is greater than the target of the gravitational force, logistics warehouse robot will be far away from the target.

(2) Artificial potential field method is used for the logistics warehousing robot to avoid obstacles. When the dynamic and static obstacles are arranged in such a way

that the logistics warehousing robot is forced to balance some special points on the way of movement, the robot cannot reach the target point.

(3) In the traditional artificial potential field method, the speed and acceleration of robot movement are not taken into account. After path planning based on the information collected around the robot by the sensor, the dynamic obstacle moves for some distance, leading to collision with the robot.

3 Principle of Fuzzy Improved Artificial Potential Field Method

3.1 Introduce Distance Adjustment Factor

According to the principle analysis of traditional artificial potential field method, the gravity of the target point of the robot will decrease with the decrease of the distance, and the repulsion of the obstacle will strengthen with the decrease of the distance. When there are obstacles near the target point, the target can not be reached. In order to solve this problem, the traditional repulsion potential field function can be improved by introducing a distance adjustment factor, which can drag the repulsion potential field to a certain extent, so that the robot's gravity is always greater than the repulsion force before reaching the target point, ensuring that the target point is the zero potential energy point in the whole driving environment. The expression of the improved repulsion potential field function is as follows:

$$U_{\text{rep}}(X) = \begin{cases} \frac{\xi}{2}\left[\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\text{lim}}}\right]^2 \cdot d^n(X_R - X_G) & d(X_R - X_O) \leq d_{\text{lim}} \\ 0 & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{6}$$

where the distance adjustment factor $d^n(X_R - X_G)$ is expressed as the relative distance between the robot and the target point, and n is the real number greater than zero.

Therefore, the corresponding expression of the improved repulsion function is as follows:

$$F_{\text{rep}} = \begin{cases} F_{\text{rep}_1(X_R,X_O)} + F_{\text{rep}_2(X_R,X_G)} & d(X_R - X_O) \leq d_{\text{lim}} \\ 0 & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{7}$$

where $F_{\text{rep}_1(X_R,X_O)}$ is the repulsion component between the robot and the obstacle, and $F_{\text{rep}_2(X_R,X_G)}$ is the repulsion component between the robot and the target point. The magnitude of the two repulsion components is as follows:

$$\left\|F_{\text{rep}_1(X_R,X_O)}\right\| = \frac{\xi}{d^2(X_R - X_O)}\left[\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\text{lim}}}\right] \cdot d^n(X_R - X_G) \tag{8}$$

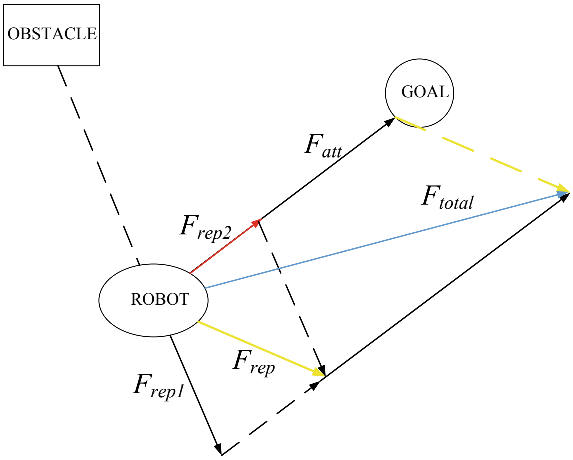


Fig. 2 Stress analysis of improved artificial potential field method

$$\left\| F_{\mathrm{rep}_2(X_R,X_G)} \right\| = -\frac{n\xi}{2}\left(\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\lim}}\right)^2 \cdot d^{n-1}(X_R - X_G) \qquad (9)$$

It can be seen that the improved repulsion is composed of two repulsion components. When the distance between the robot and the target point decreases gradually, the repulsion force will also decrease until it approaches to 0, on the contrary, the repulsion force will gradually increase to help the robot avoid obstacles and reach the target point, thus solving the problem of the target unreachable. The stress analysis diagram of the improved artificial potential field method is shown in Fig. 2.

3.2 Introduce Relative Velocity and Relative Acceleration

The traditional artificial potential field method assumes that the obstacles are regular in shape and in static state, but in the actual dynamic environment, the moving obstacles make the robot have complexity and uncertainty in the process of local path planning. Using the traditional artificial potential field method will lead to the robot unable to avoid obstacles in time and the collision phenomenon will occur. In order to solve this problem, the relative velocity and acceleration between obstacle and robot are considered. When the relative speed of the two is in the same direction and the same, the logistics storage robot does not need to avoid the obstacle when making a decision; if the relative speed of the two is reverse and relative motion, when the speed of the two is large, the robot should make a decision to avoid the obstacle, increase the repulsion effect of the obstacle on the robot, and avoid the collision between the robot and the moving obstacle. The acceleration is the trend of object movement. The introduction of the relative acceleration of robot and obstacle into the repulsion function can improve the predictability and timeliness of path planning, predict in time before the speed of obstacle changes greatly, and achieve better obstacle avoidance effect. Based on the above considerations, the relative velocity vector and relative acceleration vector are introduced on the basis of the above improved repulsion potential field function, so the new repulsion potential field function is as follows:

$$U_{\text{rep}} = \begin{cases} \frac{\xi}{2}\left[\frac{1}{d(X_R - X_O)} - \frac{1}{d_{\text{lim}}}\right]^2 \cdot d^n(X_R - X_G) + \varphi_v v_{\text{RO}} + \delta_a a_{\text{RO}} & d(X_R - X_O) \leq d_{\text{lim}} \\ 0 & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{10}$$

where v_{RO} and a_{RO} are respectively expressed as relative velocity and relative acceleration; φ_v as positive proportional gain coefficient of relative velocity; δ_a as positive proportional gain coefficient of relative acceleration.

v_{RO} refers to the component of the relative speed between the robot and the obstacle on the line between them, and a_{RO} refers to the component of the relative acceleration between the robot and the obstacle on the line between them. The specific expression is as follows:

$$v_{\text{RO}} = (v_R - v_O)^T e_{\text{RO}} \tag{11}$$

$$a_{\text{RO}} = (a_R - a_O)^T e_{\text{RO}} \tag{12}$$

where e_{RO} is expressed as unit vector, whose direction is from robot to obstacle.

The negative gradient of the new repulsion potential field function is solved, and the expression of the new repulsion function is as follows:

$$F_{\text{rep}} = \begin{cases} F_{\text{rep}_1(X_R, X_O)} + F_{\text{rep}_2(X_R, X_G)} + \varphi_v(v_R - v_O) + \delta_a(a_R - a_O) & d(X_R - X_O) \leq d_{\text{lim}} \\ \varphi_v(v_R - v_O) + \delta_a(a_R - a_O) & d(X_R - X_O) > d_{\text{lim}} \end{cases} \tag{13}$$

3.3 Fuzzy Adjustment of Repulsion Potential Field Coefficient

Through the analysis of the improved artificial potential field method, we can see that the coefficients in the repulsion potential field function are fixed. When faced with the complex and dynamic environment, the distribution of obstacles is random and uncertain. The robot can not make local obstacle avoidance behavior effectively in real time, and there is still a risk of collision. Fuzzy control algorithm has a great advantage in dealing with uncertain information, and it can deal with complex environment well. It combines the above-mentioned improved artificial potential field method and fuzzy logic control algorithm. The repulsive potential field coefficient synthetically decides the motion control parameters according to the fuzzy rules, changes the human force on the robot and then controls the change of robot motion in real time.

There are three coefficient components in the repulsive potential field function, so three fuzzy controllers with two inputs and one output are needed. The controller 1 takes the distance D_{RO} between the robot and the obstacle and its the direction angle φ as the input, the coefficient affected by the distance ξ as the output. The controller 2 takes between the relative speed V_{RO} between the robot and the obstacle and its the direction angle φ as the input, the positive proportional gain coefficient of the relative speed φ_v as the output. The controller 3 takes the distance D_{RO} between the robot and the obstacle and the distance D_{RG} between the robot and the target point as the input, and the positive proportion coefficient of the relative acceleration δ_a as the output.

Because the image of Gaussian function has no zero point, and has good smoothness and symmetry, it can describe the relationship between input and output well, so the membership function uses Gaussian function. The range of distance between robot and obstacle is (0, 1), fuzzy subset: {ZD, SD, MD, BD} = {very close, close, moderate, far} (Z: zero, S: small, M: middle, B: big); the range of distance between robot and target point is (0, 2), fuzzy subset: {ZD, SD, MD, BD} = {very close, close, moderate, far}; the domain defining the angle between the robot and the moving direction of the obstacle is (-π, π), fuzzy subset: {Nb, NM, NS, Z, PS, PM, PB} = {negative large, negative medium, negative small, zero, positive small, positive medium, positive large} (N: negative, H: tiger, P: positive). The domain defining the relative velocity of robot and obstacle is $(-1, 1)$, fuzzy subset: {NBV, NMV, NSV, ZV, PSV, PMV, PBV} = {negative large, negative medium, negative small, zero, positive small, positive medium, positive large}; the universe affected by distance, relative velocity and relative acceleration is (1, 100), fuzzy subset: {ZC, SC, MC, BC} = {gain is zero, gain is small, gain is moderate, gain is large} (N: none, C: coefficient). The fuzzy control rules of the three fuzzy controllers are shown in Tables 1, 2 and 3.

According to the completed fuzzy control rule table, the mobile robot can take the actual environment information collected as input data in the process of motion, and get different output in real time, so as to adjust the size of three coefficients in the repulsion function in real time, and play a real-time control of the repulsion force of the robot. The input and output surfaces of the designed fuzzy controller 1, fuzzy controller 2 and fuzzy controller 3 are shown in Figs. 3, 4 and 5.

Table 1 Fuzzy control rules of fuzzy controller 1

D_{RO}	φ						
	NB	NM	NS	Z	PS	PM	PB
ZD	SC	SC	MC	BC	ZC	SC	SC
SD	ZC	ZC	SC	BC	MC	SC	ZC
MD	ZC	ZC	SC	MC	SC	ZC	ZC
BD	ZC	SC	SC	SC	SC	ZC	ZC

Table 2 Fuzzy control rules of fuzzy controller 2

V_{RO}	φ						
	NB	NM	NS	Z	PS	PM	PB
NBV	BC	MC	SC	ZC	ZC	SC	MC
NMV	MC	SC	ZC	ZC	ZC	SC	SC
NSV	SC	ZC	ZC	ZC	ZC	ZC	ZC
ZV	ZC	ZC	ZC	ZC	ZC	ZC	ZC
PSV	ZC	ZC	SC	SC	SC	ZC	ZC
PMV	ZC	ZC	SC	MC	SC	ZC	ZC
PBV	ZC	ZC	MC	BC	MC	ZC	ZC

Table 3 Fuzzy control rules of fuzzy controller 3

D_{RO}	D_{RG}			
	ZD	SD	MD	BD
ZD	ZC	ZC	MC	BC
SD	ZC	ZC	MC	BC
MD	ZC	ZC	ZC	MC
BD	ZC	ZC	ZC	MC

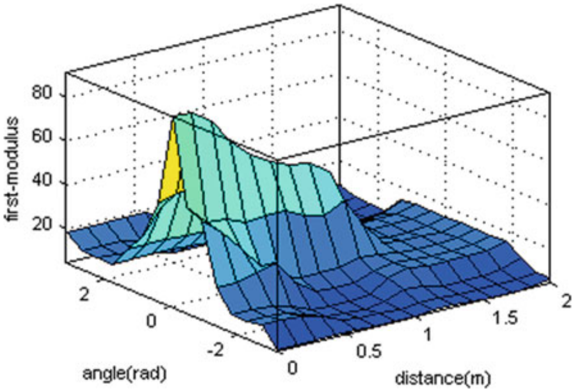


Fig. 3 Input and output surface graph of fuzzy controller 1

4 Simulation Experiment and Result Analysis

In order to verify the correctness and feasibility of the fuzzy improved artificial potential field method proposed in this paper, the static obstacle environment and dynamic obstacle environment are constructed by using MATLAB software, and the

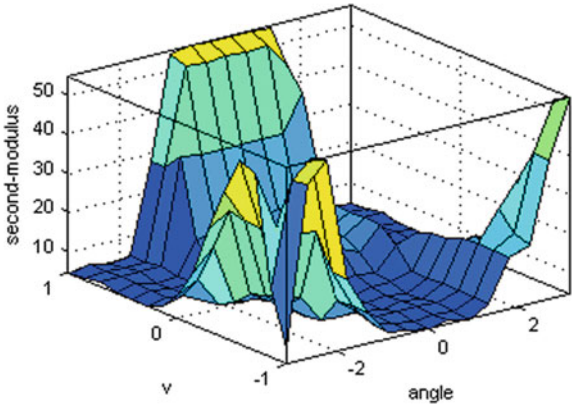


Fig. 4 Input and output surface graph of fuzzy controller 2

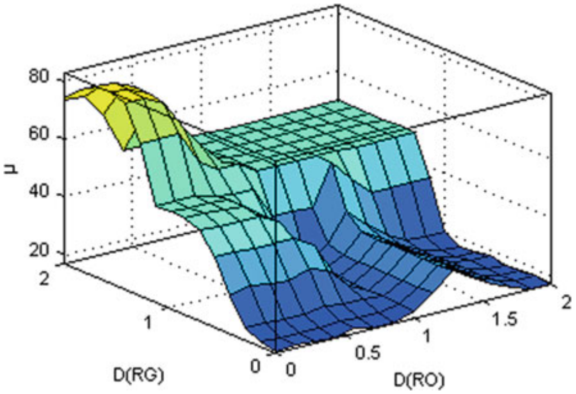


Fig. 5 Input and output surface graph of fuzzy controller 3

traditional artificial potential field method and the fuzzy improved artificial potential field method are respectively used for road strength planning simulation experiment.

(1) *When the obstacle is near the target point, the traditional artificial potential field method will lead to the robot unable to reach the target point.* In the static obstacle simulation environment, 14 static obstacle points with known position information are set up, which are represented by blue circle. The coordinates of the starting point and the target point of the logistics storage robot are (0, 0) and (10, 10). The traditional artificial potential field method and the fuzzy improved artificial potential field method are respectively used for path planning. The simulation results are shown in Figs. 6 and 7.

According to the simulation results, it can be seen that the fuzzy improved artificial potential field method can make the logistics storage robot effectively avoid obstacles

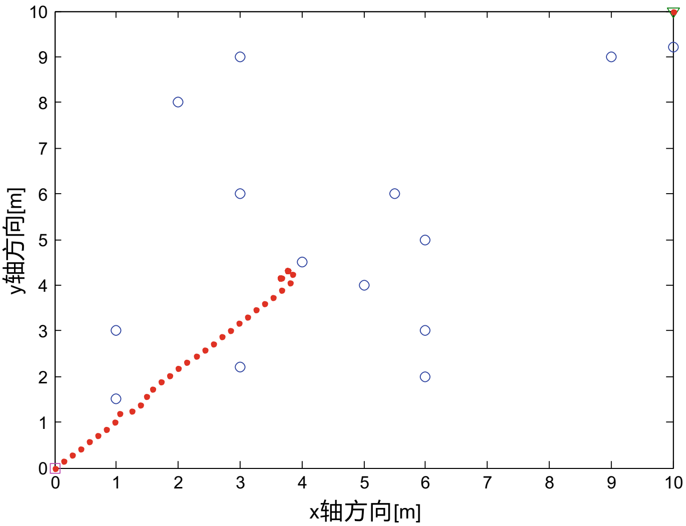


Fig. 6 Path planning of traditional artificial potential field in static environment

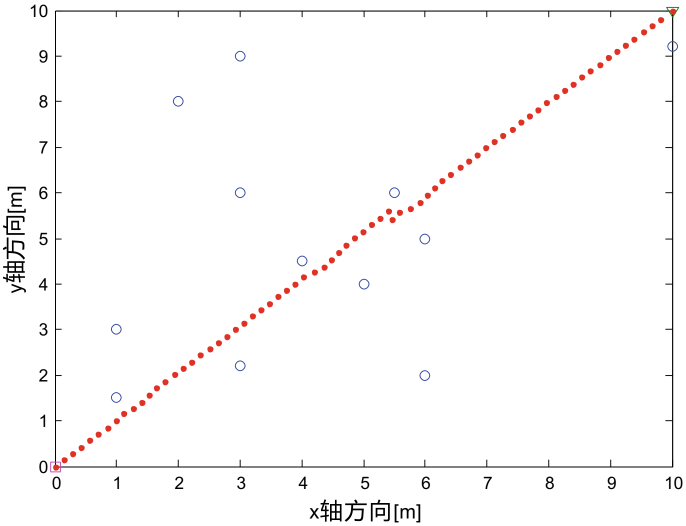


Fig. 7 Path planning of fuzzy improved artificial potential field in static environment

in the static known environment, and successfully reach the target point, which solves the problem of the traditional artificial potential field method in path planning.

(2) *The path planning simulation of mobile robot in dynamic and complex logistics storage environment.* Six static obstacles with known location information and three dynamic obstacles with known movement information are set in the logistics storage

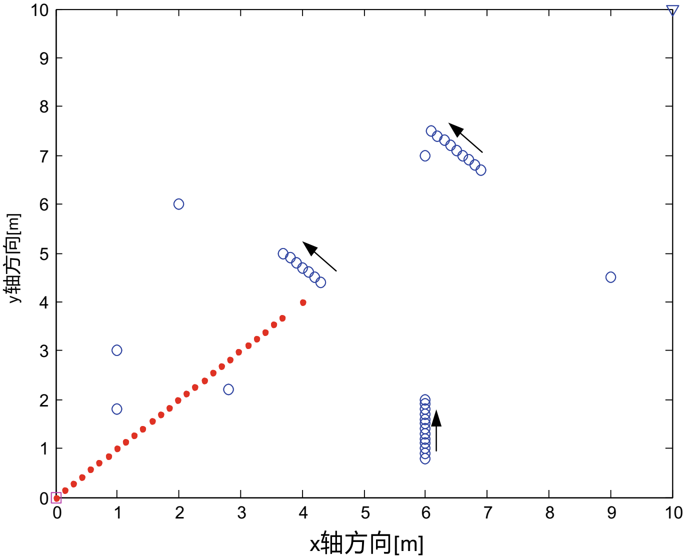


Fig. 8 Path planning of traditional artificial potential field in dynamic environment

environment, among which a single blue circle represents static obstacles, multiple groups of blue circles constitute dynamic obstacles, and the black arrow represents the movement direction of dynamic obstacles. The starting point coordinate of the mobile robot is (0, 0) and the target point coordinate is (10, 10). The traditional artificial potential field method and the fuzzy improved artificial potential field method are respectively used for path planning. The simulation results are shown in Figs. 8 and 9.

It can be seen from the simulation results that when encountering dynamic obstacles, the traditional artificial potential field method can not help the mobile robot to make effective obstacle avoidance behavior, resulting in the robot unable to reach the target point; while the fuzzy improved artificial potential field method can effectively help the mobile robot to avoid static obstacles and dynamic obstacles in real time, and finally complete the path planning and successfully reach the target point.

5 Conclusions

In order to improve the ability of local path planning of mobile robot in logistics storage, a fuzzy improved artificial potential field method is proposed in this paper. In order to solve the problem of inaccessibility of the traditional artificial potential field method, a distance adjusting factor is introduced to improve the original repulsion potential field function, so that the force on the robot before reaching the target point is not zero, and the target point is the global minimum point. At the same time,

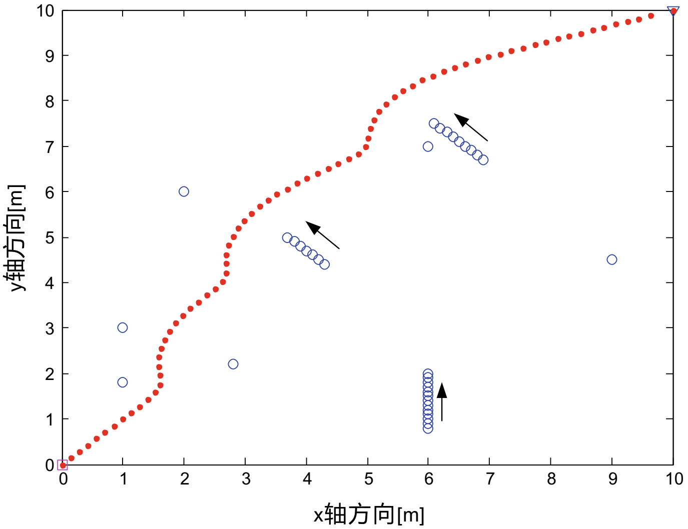


Fig. 9 Path planning of fuzzy improved artificial potential field in dynamic environment

in order to enable the robot to deal with the complex and dynamic logistics storage environment, the relative speed and acceleration of the robot and the obstacles are introduced. Combined with the fuzzy logic control algorithm, the three coefficients of the improved repulsion function are adjusted in real time, and the comprehensive decision is made to achieve the high-efficiency obstacle avoidance. Finally, the MATLAB simulation experiment is carried out. The experimental results show that in the static or dynamic environment, the fuzzy improved artificial potential field method can make the robot avoid obstacles in time and quickly, plan a smooth path and reach the target point smoothly, which proves the effectiveness and feasibility of the algorithm.

Acknowledgements This work was supported by the Scientific Research Program of Beijing Education Commission under Grant KM201910015004.

References

1. S.T. Zhao, X. Jia, Research progress and trend of intelligent manufacturing and its core information equipment. Mech. Sci. Technol. Aerosp. Eng. **36**(1), 1–16 (2017)
2. W.R. Shi, X.H. Huang, W. Zhou, Path planning for mobile robots based on improved artificial potential field method. J. Comput. Appl. **30**(8), 2021–2023 (2010)
3. Z.Y. Feng, Soil suitability evaluation based on bp artificial neural network: the case of Jiangxia district. Environ. Eng. Manage. J. **17**(1), 229–235 (2018)

4. D. Wu, Y. Liu, K. Zhou, K. Li, J. Li, A multi-objective particle swarm optimization algorithm based on human social behavior for environmental economic dispatch problems. Environ. Eng. Manage. J. **18**(7), 1599–1607 (2019)
5. O. Khatib, Real-time obstacle avoidance for manipulators and mobile robots. Int. J. Robot. Res. **5**(1), 90–98 (1986)
6. H.S. Min, Y.H. Lin, S.J. Wang, Path planning of mobile robot by mixing experience with modified artificial potential field method. Adv. Mech. Eng. **7**(12), 1–17 (2015)
7. S. Fazli, L. Kleeman, Wall following and obstacle avoidance results from a multi-DSP sonar ring on a mobile robot, in *Niagara Falls: Proceedings of the IEEE International Conference on Mechatronics and Automation* (2005), pp. 432–437
8. D.C. Chen, S. Li, L.F. Liao, A recurrent neural network applied to optimal motion control of mobile robots with physical constraints. Appl. Soft Comput. J. 85 (2019)
9. Z.K. Ying, Flight path planning of agriculture UAV based on improved artificial potential field method, in *The 30th China Conference on Control and Decision-making*
10. T. Liu, H.B. Li, Z.X. Duan, Path planning for mobile robots based on artificial force field. Comput. Simul. **197**(11), 144–146 (2007)
11. F.C. Huo, J. Chi, Z.J. Huang, L. Ren, Q.J. Sun, J.L. Chen, Overview of path planning algorithms for mobile robots. J. Jilin Univ. (Inf. Sci. Edn.) **36**(06), 639–647 (2008)
12. P.B. Zhang, Y.P. Zeng, Path planning for mobile robots based on improved artificial potential field method. Robot. Technol. Appl. **3**, 27–29 (2018)
13. D.F. Li, H.B. Deng, Z.H. Pan, T. Peng, C. Wang, Trajectory tracking control law of snake robot avoiding obstacles in flow field based on improved snake curve. Robot **41**(04), 433–442
14. R. Meng, W.J. Su, X.F. Lian, Path planning for mobile robots based on dynamic fuzzy artificial potential field method. Comput. Eng. Des. **31**(07), 1558–1561 (2010)
15. J.Y. Liu, Research on obstacle avoidance algorithm of mobile robot based on artificial potential field method. Ph.D. dissertation, Central China Normal University, Wuhan, HB, China (2018)
16. W. Han, K.B. Sun, Intelligent omnidirectional vehicle path planning based on fuzzy artificial potential field method. Comput. Eng. Appl. **54**(06), 105–109 (2008)
17. L.A. Ocampo, C.M. Himang, A. Kumar, M. Brezocnik, A novel multiple criteria decision-making approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy AHP for mapping collection and distribution centers in reverse logistics. Adv. Prod. Eng. Manage. **14**(3), 297–322 (2019)
18. M.W. Shen, Y.J. Yu, D.J. Yan, F.W. Hai, L.H. Bao, A new positioning method based on multiple ultrasonic sensors for autonomous mobile robot. Sensors (Basel, Switzerland) (1), 20 (2019)
19. G.I. Fragapane, C. Zhang, F. Sgarbossa, J.O. Strandhagen, An agent-based simulation approach to model hospital logistics. Int. J. Simul. Modelling **18**(4), 654–665 (2019)

The Whole-Course Logistics Enterprise Credit Reference System Based on Blockchain Technology

Yang Mao, Minzhen Huang, Shifeng Liu, Yi Song, Guohua Li, and Jingya Liu

Abstract In order to solve the problem of credit information asymmetry between whole-course logistics enterprises, this paper proposes a credit reference system structure based on the blockchain technology and expounds the application processes of data collection, credit modeling scoring and credit score sharing by blockchain technology. Based on the data of A-share listed logistic companies in China from 2011 to 2019, this paper empirically examines the scoring model and discusses the advantages of blockchain application in the logistics credit industry in the era of big data: dynamic update and traceability of credit data, promoting credit data sharing, ensuring the privacy of information subjects and unifying industry credit evaluation standards.

Keywords Blockchain · Whole-course logistics · Scoring model · Credit reference system

Y. Mao · S. Liu · Y. Song (✉) · J. Liu
Department of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 19113051@bjtu.edu.cn

Y. Mao
e-mail: 18113060@bjtu.edu.cn

S. Liu
e-mail: shfliu@bjtu.edu.cn

J. Liu
e-mail: 19120615@bjtu.edu.cn

M. Huang · G. Li
Institute of Electronic Computing Technology, China Academy of Railway Sciences, Beijing, China
e-mail: huangminzhen321@163.com

G. Li
e-mail: guohualea@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_20

1 Introduction

Whole-course logistics is a new type of logistics mode. Its basic meaning is to achieve the effect of improving logistics efficiency and reducing logistics costs through the cooperation of different enterprises in logistics [1]. The key point is to unify multiple transportation modes. In the process of logistics cooperation between enterprises, credit information plays an important role, and the asymmetry of enterprise credit information is one of the difficulties faced by whole-course logistics. As a new application mode of distributed data storage, point-to-point transmission, consensus mechanism, encryption algorithm, and other computer technologies, blockchain technology has the characteristics of decentralization, traceability, immutability, cross-platform, distributed sharing et al. [2], and has now gradually been applied in many fields, such as finance, public services, digital rights, insurance, public welfare et al. As the cutting-edge modern information technology, blockchain technology is gradually being tried in the field of logistics, but most of the researches focus on logistics supply chain [3, 4], lack of comprehensive integration research on blockchain, logistics, and credit. The whole-course logistics credit reference system based on blockchain technology is crucial to reduce the asymmetry of enterprise credit information in the whole-course logistics and improve business efficiency. The application of blockchain technology should be promoted in the field of logistics credit reference, and the logistics industry should also grasp the future development trend of this key technology, which will further promote the development of the entire logistics industry.

2 Literature Review

The main reason for the credit problem of whole-course logistics enterprises in is the lack of effective credit reference system. How to evaluate the credit of logistics enterprises has always been the concern of scholars. Guo analyzed the current situation of China’s third-party logistics credit evaluation market, sorted out the existing problems in China’s logistics credit evaluation market, and put forward suggestions [5]. Wu and Yang believed that the current credit problems in the logistics market are prominent. How to select the enterprises with good credit conditions and how to make analysis and early warning of the enterprise credit conditions have become urgent problems to be solved. They analyze the importance of credit problem to the long-term stable development of logistics enterprises and puts forward the establishment of credit supervision system which can carry out an objective evaluation in combination with the causes of the lack of credit of logistics enterprises in the current market environment. They put forward a credit evaluation index system based on quantitative analysis and finally the fuzzy neural network algorithm is determined as a method to evaluate the credit status of logistics enterprises, which is prepared for further quantitative analysis with data [6]. Cao and Lu conducted an empirical test on the ability of the Z-score Model to predict the credit status of logistics listed companies. The research conclusion is that it has a more accurate prediction ability, which proves that the financial data of logistics companies is effective. Therefore, analysts can use the financial data of logistics companies to predict the credit status of companies in a certain period [7]. Miao et al. used the method of extracting principal components to build a logical regression model and original indicators to build a logical regression model to carry out financial early warning research of logistics enterprises and came to the following conclusion: using the logical regression model based on clustering analysis method to judge whether the superior company is in financial difficulties can reduce errors and is more practical [8]. Hu et al. used a logistic regression model to build a financial crisis early warning model of logistics listed companies and got good prediction results [9].

Although blockchain technology originated from digital currency management, the application of blockchain technology has gone far beyond the scope of digital currency management, and gradually extended to various market transactions, such as logistics supply chain, insurance, industry, intellectual property, and credit reference system construction [10, 11]. Byström discussed how blockchains potentially could affect the way credit risk is modeled, and how the improved trust and timing associated with blockchain-enabled real-time accounting could improve default prediction [12]. Liu et al. built a multi-source credit tracing system for small and medium-sized logistics enterprises based on blockchain technology [13]. Wang proposed the framework of a personal credit system of multicenter and distributed alliance blockchain and pointed out the development path of establishing a personal credit system in the Internet era through "one library and one channel" [14]. Guo and Song proposed a credit system structure based on blockchain technology and two data transaction patterns and the technical framework of the credit data trading platform [15]. Wang and Zhang believed that in the information era, blockchain technology will have a profound impact on the construction of the social credit system, and one of the significant changes is to reduce the cost of trust [16]. Ju et al. incorporated the blockchain technology to designed a multi-source data sharing framework to supporting future credit system, and based on the multi-source data sharing of the blockchain, a big data credit platform for multi-source heterogeneous data fusion was established using artificial intelligence, data mining, intelligent contract and other methods [17]. Che elaborated the feasibility of the application of blockchain in the credit industry, listed the application cases of blockchain in the credit industry at home and abroad, and introduced the application scheme of blockchain in the credit industry in China [18]. Ta and Li established a cross-platform credit data-sharing model based on blockchain and discussed its mechanism and application scenarios. The empirical results show that: the number of institutions in the credit system and overdue rate shows a significant negative relationship. After the blockchain technology is used to open up the credit platform, the number of institutions providing credit services in the system can be greatly increased [19]. Shi believed that as a bottom technology and development vision, blockchain is gradually expanding in the field of credit information services. Based on the technical characteristics, the application of blockchain to credit information services can meet the partial demand of economic development for the social credit system construction and have congenital application advantages [20]. Pan studied the application of blockchain technology in big data credit and believed that blockchain combined with big data credit reference can effectively solve various problems in credit reference [21]. Chen expounded the principle and technical points of blockchain and its feasibility in the field of credit reference and pointed out the limitations of blockchain technology in private key loss, user forgotten privilege, system security, and information storage [22].

3 Whole-Course Logistics Enterprise Credit Scoring Modeling

This section is divided into three parts. In the first part, we firstly carry out credit portrait for whole-course logistics enterprises, establishes an index system while the second part describes the whole-course logistics enterprise credit-scoring model. An empirical study on the credit rating model of the whole-course logistics enterprises is introduced in the third part.

3.1 Index System

Based on the summary of previous research literature on credit risk early warning of logistics enterprises [6–9], this paper selects indicators with high frequency, including 18 indicators in six aspects, namely profitability, growth ability, operation ability, solvency, cash flow, and judicial risk (see Table 1).

3.2 Credit Scoring Model

The ideal credit reference model of whole-course logistics enterprises should be able to input the characteristics of the whole-course logistics enterprises' credit and predict the category of the whole-course logistics enterprises, that is, whether the whole-course logistics enterprises will default, which is the main objective of statistical inference on the credit scoring problem. The logistic regression model has widely used for analyzing the relationship between a group of independent variables and discrete dependent [23]. Discrete dependent variables are defined as 0, 1, 2 et al. Discrete values of variables, for example, the application scenario of the model in the whole-course logistics enterprises credit evaluation, the independent variable x is the feature of all dimensions of the whole-course logistics enterprises credit evaluation, and the dependent variable y is the binary variable with the value of 0 and 1 respectively when the whole-course logistics related enterprises breached the

Table 1 Index system

Level 1 index	Level 2 index	Division standard
Profitability	Return on equity	Net income/shareholders' equity
	Return on assets	Net profit/average total assets
	Net profit margin	Net profit/total operating income
Solvency	Debt to asset ratio	Total liabilities/total assets
	Quick ratio	Quick assets/current liabilities
	Current ratio	Current assets/current liabilities
Operation ability	Total asset turnover	Net operating income/total average assets
	Current asset turnover	Net income from main business/total average current assets
	Receivables turnover ratio	Net credit sales income/average balance of accounts receivable
	Fixed asset turnover	Net income/average net value of fixed assets
Growth ability	Growth rate of net profit	Current net profit/base period net profit
	Growth rate of total assets	Current total assets/base period total assets
	Growth rate of operating income	Current net assets/base period net assets
Cash flow	Cash flow income ratio	Net cash flow/operating income
	Total cash debt ratio	Net cash flow/total debt
	Cash flow ratio	Net cash flow/current liabilities
Judicial risk	Dishonest execution	The executed enterprise has the ability to perform but fails to perform the obligations determined in the effective legal document

contract, it is recorded as y = 1, and if there is no serious breach of contract, it is recorded as y = 0.

A vector with m independent vectors is $x = (x_1, x_2, x_3, \ldots, x_m)$, and let the conditional probability $P(y = 1|x)$ be the probability of an event according to the observed quantity. The Logistic model can be expressed as:

$$P(y = 1|x) = \pi(x) = 1/\left(1 + e^{-g(x)}\right) \tag{1}$$

$$g(x) = \beta_{0+}\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m \tag{2}$$

Under the x condition, the probability that y does not occur (that is, y = 0) is:

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - 1/\left(1 + e^{-g(x)}\right) = 1/\left(1 + e^{g(x)}\right) \tag{3}$$

Therefore, the ratio of the probability of the whole-course logistics enterprises' default and non-default is:

$$P(y = 1|x)/P(y = 0|x) = p/(1 - p) = e^{g(x)} \tag{4}$$

This ratio is denoted as the advantage ratio, which can actually be considered as the occurrence ratio of whole-course logistics enterprises' default and non-default. If the natural logarithm is taken, the following can be obtained:

$$\ln\left(\frac{p}{1 - p}\right) = \beta_{0+}\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m \tag{5}$$

The goal of the logistic model is to solve this set of weights $\beta_1, \beta_2, \ldots, \beta_m$, and the estimation can be obtained by maximum likelihood estimation.

The output of the logistic regression model is the logarithm of the ratio of the probability of the whole-course logistics enterprise's default and non-default, which can be converted to the credit scoring. Linear transformation of the logarithm, followed by a constant value, makes the final credit score of whole-course logistics enterprises fall within a predetermined score range. The higher the credit score, the better the credit of whole-course logistics enterprises. The transformation formula is:

$$Score = \ln(P/(1 - P)) \times \alpha + \beta \tag{6}$$

In the equation, P represents the probability of the whole-course logistics enterprises' default, while is the coefficient of the linear transformation. Generally speaking, it includes a logarithm value, so that the final credit score of the whole-process logistics enterprises is within the predetermined score range.

3.3 Empirical Study

The research on the credit risk of listed companies usually takes special treatment (ST) due to abnormal financial situations as a sign of credit risk [24]. We extracted our sample from the ST logistic companies and non-ST logistic companies actively listed on Shanghai and Shenzhen Stock Exchange Market from 2011 to 2019. Our final sample consists of 79 companies, there are 7 ST companies and 72 non-ST companies. They were located in five industries of whole-course logistics: Handling and transportation agent industry, railway transportation industry, road transportation industry, shipping transportation industry, and warehousing industry.

We take these logistic listed corporations as the research sample to testify the validity of the index system and credit-scoring model. Data were extracted from Choice Database.

Because there are few negative samples and the samples are obviously unbalanced, the total amount of negative samples is increased by oversampling. In this paper,

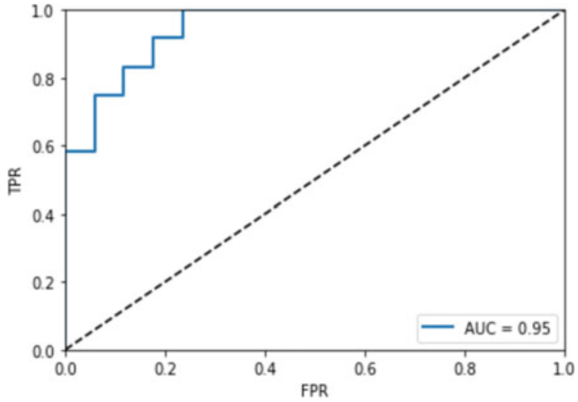


Fig. 1 Receiver operating characteristic curve

the smote algorithm is used to realize the oversampling of negative samples. After processing, 80% of the samples are randomly selected as the training set, 20% of the samples are used as the test set. After training the model with the training set, the performance of the model is tested with the test set.

Figure 1 shows the prediction results outside the sample after using the logistic regression method to build the credit risk model, AUC = 0.95. The empirical study results show that the model is excellent for whole-process logistics enterprises' credit classification.

4 Whole-Course Logistics Enterprise Credit Reference System Based on Blockchain Technology

Regarding the problems in the credit reference of whole-course logistics enterprises, a credit reference framework based on the alliance chain-based blockchain architecture can be adopted to reshape the entire architecture, as shown in Fig. 2.

In the new credit reference system, the modeling node, the railway transportation enterprises, the highway transport enterprises, the shipping enterprises, the warehousing enterprises, the handling and transportation agent in the whole-course logistics transportation chain are participants, first of all, all the whole-course logistics enterprises extract and collect their own credit data in real-time according to the index system and package them as credit data sets. After being encrypted and verified, these credit data sets are called up and recorded in the blockchain database. With smart contract technology, the credit scoring model is written into the model node of the blockchain system. The real-time credit data of the enterprise establishes a scoring model to provide corresponding credit scoring for all whole-course logistics enterprises. Credit scores can only be recorded in the blockchain after being verified by broadcasting on the chain, while protecting the privacy of the whole-course logistics enterprises, they can also obtain transparent credit index, which can more fully

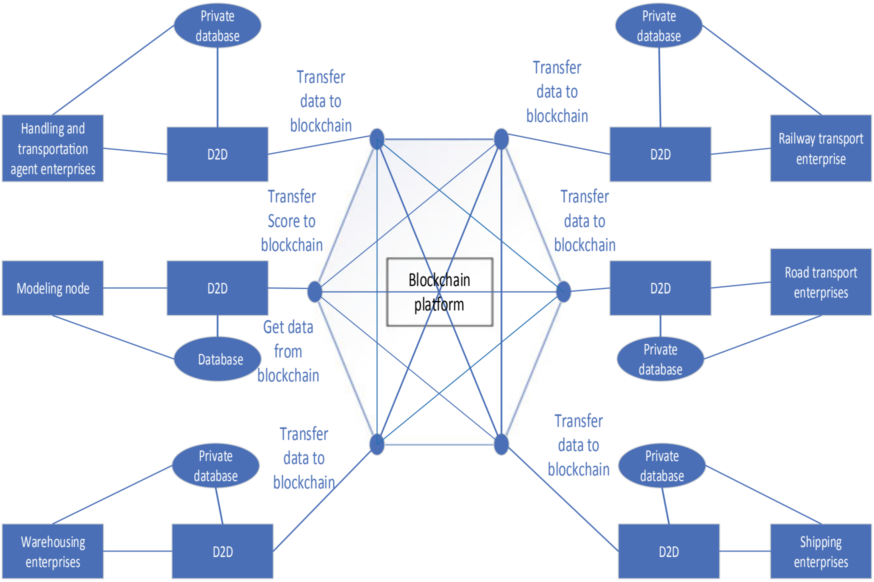

Private database
Handling and transportation agent enterprises
D2D
Transfer data to blockchain
Transfer data to blockchain
Private database
D2D
Railway transport enterprise
Transfer Score to blockchain
Transfer data to blockchain
Modeling node
D2D
Blockchain platform
D2D
Road transport enterprises
Database
Get data from blockchain
Private database
Transfer data to blockchain
Transfer data to blockchain
Private database
Private database
Warehousing enterprises
D2D
D2D
Shipping enterprises

Fig. 2 Whole-course logistics enterprises credit reference system based on blockchain technology

evaluate the credit of the whole-course logistics enterprises, strengthen the whole logistics participation in the circulation of credit data among enterprises, in order to stimulate the vitality of the logistics market.

Blockchain credit reference system is carrying the whole-course logistics chain of each enterprise credit data and credit scores, which constitute a whole-course logistics enterprise' credit information data chain that cannot be tampered with. In this way, it can realize the digitalization of the relevant credit information assets of enterprises involved in the whole-course logistics, and build a decentralized and fair cooperation trust platform based on the blockchain technology. The credit system can reduce the credit risk and cost of major whole-course logistics participants such as handling and transportation agent enterprises, railway transport enterprise, road transport enterprises, shipping enterprises, and warehousing enterprises in the process of collaboration, open up the trading relationship between upstream and downstream entities. The credit of core enterprise in the trading main body transfer to no direct transaction's enterprise, so as to solve the distal enterprise credit information asymmetry problem, to ensure that each credit body accumulation of credit data is reliable, not been tampered with. The credit scores of the whole-course logistics enterprises based on these credit data can be used as the basis for the credit risk assessment of both parties in subsequent transactions. The specific process is as in Fig. 3.

The whole-course logistics enterprise on the alliance chain first uploads the data to the blockchain. After the consent of the participating nodes on the alliance chain, the model node reads the data from the blockchain according to the smart contract, then, node modeling and scoring based on data from blockchain and list of dishonest

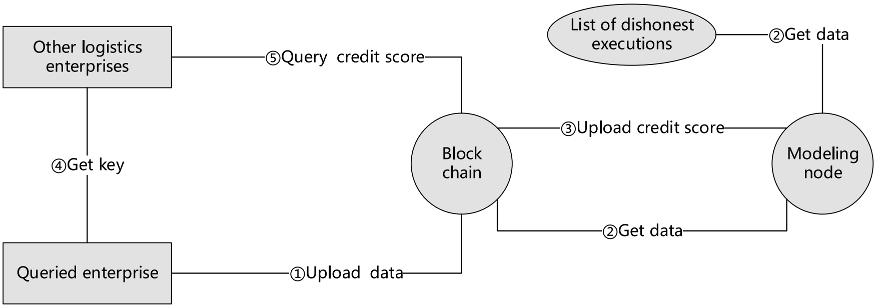


Fig. 3 The specific process of credit reference based on blockchain technology

executions obtained from the court. The modeling node uploads the credit score to the blockchain. When other logistics enterprises need to query the credit score of a logistics enterprise, they can obtain the key from the queried enterprise. Then the key can be used to obtain the credit score of the logistics enterprise from the blockchain, as shown in Fig. 3.

5 The Advantages of Whole-Course Logistics Enterprise Credit Reference Systems Based on Blockchain Technology

Compared with the traditional credit reference system, blockchain technology is decentralized, immutable, traceable, and smart contracts, making it suitable as a "credit lubricant" for the credit reference. Therefore, it is convenient to combine it with credit information of the whole-course logistics enterprise. Specifically, the whole-course logistics enterprise credit reference system based on blockchain technology built in this article has the following advantages.

5.1 Eliminate the Credit Information Barriers

Blockchain technology is the cornerstone of credit in whole-course logistics. If there are multiple parties involved, it can provide immutable information storage and retrospective query, and provide participants with authentic, credible, and non-tamperable credit on-chain data and credit scores. Each participant can obtain credit data from each other based on blockchain technology endorsement. In this way, the misgivings related to the enterprises in the whole-course logistics of credit data exchange are eliminated, and the credit information barriers between multimodal transport participants during the whole-course logistics are eliminated, thereby reducing the credit cost of whole-course logistics and improves the efficiency of whole-course logistics.

5.2 Ensuring Data Security

During the process of data collecting, modeling, and scoring, the blockchain's encryption technology is used to encrypt and desensitize the enterprise's credit data without revealing key information of the enterprise. The digital signature generated by the hash function makes the data difficult to tamper with. During the data transmission process, the encrypted file is generated by the public key. The private key of the data is only owned by the data owner. It is guaranteed that the source data collected on the chain can only be controlled by the enterprise itself, and other participating enterprises can only view the credit score of the enterprise. Therefore, the data of the whole-course logistics enterprise have achieved security.

5.3 Dynamic Update of Credit Information

Blockchain technology will update, collect and store the credit data of whole-course logistics enterprise in real-time, and provide real-time and dynamic data for the credit modeling scores of whole-course logistics enterprise. With the development of credit reference systems for whole-course logistics enterprises, new enterprises are constantly joining the logistics area. Multiple logistics and transportation chains are intertwined with each other, forming a whole-course logistics credit network, and finally converging into a whole-course logistics dynamic enterprise credit database, further improving the credit scoring model in order to dynamically update whole-course logistics enterprise credit scoring in real-time.

5.4 Credit Information Can Be Traced

Based on the decentralization of blockchain technology, it does not rely on a centralized node, and uses trusted technology to publicly record all information in the "public ledger". The data on the chain is time-stamped and cannot be tampered with. The credit reference framework further formed multi-source cross-validation. whole-course logistics enterprises are linked from the generation of data, making the data on the chain immutable and traceable, which is conducive to ensuring the authenticity of the credit information of the whole-course logistics enterprise.

5.5 Forming Consensus Credit Evaluation

Through smart contracts, an enterprise credit evaluation system with multiple parties participating can be embedded in the blockchain in the form of code. Due to the dynamic nature of credit data and the autonomy of the blockchain, the system can automatically identify the dynamic credit trends of whole-course logistics enterprises without human intervention. In this way, the credit score can be adjusted in real-time and dynamically to form a practical credit evaluation consensus, which will eventually promote the logistics industry to establish credit evaluation standards.

6 Conclusions

This article discusses the unsolved credit problems of the whole-course logistics enterprise, proposes to use blockchain technology to solve them, and establishes a credit reference system based on blockchain technology. This system can eliminate the barriers to credit information between enterprises, protect corporate data security, dynamically update credit information, ensure the authenticity of credit information, and traceback credit data. This system transforms the traditional centralized credit reference system into a decentralized, difficult to tamper with, and traceable dynamic credit reference system. The new credit reference system integrates the blockchain technology and the credit scoring model, which make warning of the whole-course logistics enterprise's credit so that the enterprise's credit information becomes more transparent as well. Under the new credit reference 'system, the untrustworthy behavior of whole-course logistics enterprise can be significantly improved, so that the problem of asymmetric credit information can be effectively solved. However, because the application research of blockchain technology in various fields has just started, further in-depth research remains to be explored. This paper provides a solution for the application of blockchain technology in the credit reference system of whole-process logistics enterprises and hopes to provide a reference for further research on the integration of blockchain and logistics credit reference.

Acknowledgements This research was done as part of the "One belt, one road service area, railway logistics and block chain integration, innovative application technology research" project (J2019X008) funded by Science and technology research and development plan of China National Railway Group Co., Ltd.

References

1. B. Zhang, Whole-course logistics management based on the value chain. China Storage Transp. **3**, 42–44 (2003)
2. L. Zhang, B. Liu, R. Zhang, B. Jiang, Y. Liu, Overview of blockchain technology. Comput. Eng. **5**, 1–12 (2019)
3. X. Zhang, Research on supply chain system construction of new logistics industry—analysis based on block chain technology. J. Tech. Econ. Manage. **7**, 103–107 (2019)
4. C. Yuan, Research on the application of block chain in logistics supply chain. Serv. Sci. Manage. **8**, 142–146 (2019)
5. Z. Guo, The third party logistics credit evaluation market problems and countermeasures. Pearl River Water Transp. **22**, 94–95 (2019)
6. C. Wu, K. Yang, Feasibility study on credit evaluation of logistic enterprises based on quantitative analysis. Logistics Sci-Tech **11**, 25–28 (2019)
7. X. Cao, Q. Lu, An empirical study on financial risk early warning of logistics listed companies in China. China Logist. Purchasing **21**, 74–75 (2010)
8. P. Miao, H. Sun, Y. Gu, Financial early-warning of listed logistics companies based on logistics model. Logist. Technol. **2**, 71–75 (2020)
9. M. Hu, Y. Ding, K. Zhang, Research on financial crisis early warning of listed logistics companies in China. China Storage Transp. **2**, 118–121 (2015)
10. Y. Tian, G. Zhao, L. Shen, Blockchain transportation: taking freight logistics and the market governance as the example. China Business and Market **2**, 50–56 (2018)
11. T. Aste, P. Tasca, T. Matteo, Blockchain technologies: the foreseeable impact on society and industry. Computer **9**, 18–28 (2017)
12. H. Byström, Blockchains, real-time accounting and the future of credit risk modeling. Ledger **4**, 40–47 (2019)
13. R. Liu W. Shen, C. Tang, On the traceable system of multi-source credit reporting for small and medium logistic companies based on blockchain technology. Credit Reference **10**, 32–36 (2019)
14. Z. Wang, An analysis on the league blockchain model of personal credit reporting system in the internet era. Credit Reference **8**, 26–32 (2019)
15. S. Guo, Z. Song, Design and application of blockchain pattern for credit information industry. Chin. J. Netw. Inf. Secur. **4**, 63–71 (2018)
16. G. Wang, J. Zhang, Discussion on blockchain-based credit reporting system in the information era. Credit Reference **9**, 13–17 (2018)
17. C. Ju, J. Zou, X. Fu, Design and application of big data credit reporting platform integrating blockchain technology. Comput. Sci. **11**, 522–526+552 (2018)
18. Y. Che, The application exploration of block chain in credit reference industry. Fin. Technol. Time **10**, 20–23 (2018)
19. L. Ta, M. Li, An analysis of the prospects for application of blockchain technology in internet financial credit. J. Northeastern Univ. (Soc. Sci.) **5**, 466–474 (2018)
20. M. Shi, Discussion on the application of block chain technology in credit reporting industry. Credit Reference **1**, 20–24 (2018)
21. F. Pan, Discussion on the Application of Block Chain Technology in Credit Reporting Industry. Economic Research Guide **29**, 159–160 (2019)
22. K. Chen, Discussion on the application of block chain technology in credit reporting industry. Econo. Res. Guide **21**, 153+166 (2019)
23. M. Bensic, N. Sarlija, M. Susac, Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intell. Syst. Account. Finance Manage. **13**, 113–150 (2005)
24. Y. Liu, Y. Chen, Tone at the top and credit risk warning for listed companies: textual analysis of company annual reports. Fin. Econ. Res. **4**, 46–54 (2018)

Predicting Question Response Rate in Different Topics in Zhihu

Hong-de Liu, Ming Li, and Cheng-qi Yang

Abstract As a popular social question answering platform in China, Zhihu is facing the problem of a low question response rate. In order to find out the reason, we conduct a predictive research on the question response rate of different topics. By comparing the influencing factors of the question response rate of different topics, we tend to find a solution which is beneficial to increase users' activity. We obtain a total of 4293 data in six hot topics. A binary logistic regression method is used to construct a question response rate prediction model for different topics from the characteristics of the question and the questioner. Using 600 data to verify 6 prediction models, the prediction accuracy rate of which is more than 80%. We find carrier richness, question length, polite expression, expression of urgency and reliability of the questioner affect the question response rate of six topics. Moreover, posting period, affective tendency expression, the questioners degree of social learning, extended centrality and inward centrality cast different effects in different topics.

Keywords Zhihu · Question response rate · Binary logistic regression · Six topics

1 Introduction

The birth of the social question answering platform (Social Q & A Platform) has changed the way of knowledge sharing and dissemination. Social Q & A Platform provides online question and answering service, where users can post their own questions in life, and answer various questions of others [1]. Popular Social Q & A Platforms include Quora, Facebook, and Twitter, which use existing relationship

H. Liu · M. Li (✉) · C. Yang
School of Economics and Management, China University of Petroleum, Beijing, Changping District 102249, China
e-mail: limingzyq@foxmail.com

H. Liu
e-mail: lhd13180308602@sina.com

C. Yang
e-mail: ycq128128@126.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_21

groups to accelerate the dissemination of information [2]. The Chinese social Q & A platform started late, and the most popular ones are Baidu Know, Zhihu, and Husk. When users encounter problems, if search engines return very poor results for their problems, they will use social Q & A platforms [3]. However, social Q & A platforms are facing the problem of low response rates [4], which reduces users' activity and loyalty to the platform.

As the most popular social Q & A platform in China, Zhihu is centered on knowledge sharing and uses social networks to establish connections between people so that high-quality knowledge can be more widely spread. Users use the platform to share knowledge, experience and insights all the time. We randomly crawled 1527 questions in Zhihu and found that only 724 of them were answered, and the response rate was less than 50%. From the data above, we find that there is a problem of low response rate in Zhihu.

2 Related Research

Foreign scholars have little research on Zhihu. At present, there are mainly following two categories of domestic scholars' research about Zhihu. One is about knowledge payment research: Ding [5], Shi [6]analyzed the current status of knowledge payment development; the other is about response research: Li and Zhang [7], Deng [8], Liu and Lin [9]studied the quality of answering questions. In short, there are currently few researches on predicting users' question response rate. For the research on the prediction of the response rate of social Q & A platforms, most scholars mainly focus on the motivation of using social Q & A platforms. As for extrinsic motivation research: Choi [10] found that the establishment of the answer wealth value will increase the enthusiasm of the respondents. Oh [11] analyzed 10 motives; with regards to intrinsic motivation research: Dearman and Truong [12] studied the impact on the response rate from the perspective of the questioners. Nam and Ackerman [13] confirmed that the driving force for user to response to questions in social Q & A platform includes learning new knowledge, helping others, or considering it as a hobby. Deng [14] took Baidu Know as a whole to predict the question response rate of the Q&A platform. To sum up, there is no related research on the prediction of question response rate of Zhihu, nor does it consider the influencing factors of question response rate under different topics.

3 What is Binary Logistic Regression

Based on a comprehensive analysis of user behavior prediction methods, we find that the question response rate is a typical binary classification problem [14], and the logistic regression model is the most commonly used multivariate quantitative analysis method for regression analysis of binary classification dependent variables

[15]. Therefore, we use binary logistic regression to predict the response rate of questions in social Q & A platform.

$$\mathrm{P}(Y = 1|X) = f(X) = \frac{1}{1 + e^{-\mathrm{g}(X)}} \tag{1}$$

P is the probability of the event of the dependent variable Y = 1, and X is the independent variable.

4 Research Models and Hypothesis

4.1 What is the Characteristics of the Question

(a) *Carrier Richness:* The content of general information will be presented to users through two forms of carriers: On the one hand, there is a kind of carrier for presenting information, such as text, sound, pictures, videos; on the other hand, there is also a physical carrier for storing and transmitting information, such as paper, film, disk, CD [16]. Previous research found abundant carrier attract more users. Putting pictures in answers can enhance the reading experience [14, 17]. So we make a hypothesis as below:
H1: Carrier richness of the question content is positively related to the question response rate.

(b) *Question Length*: Questions raised in Zhihu are limited to a maximum of 50 characters (including punctuation). Researchers has shown that long questions are often more difficult to read and understand than short questions. To save time and effort, people who answer questions may avoid answering lengthy questions. Questions with an average length of 5–10 words often get more answers than other questions [18]. So we make a hypothesis:
H2: The length of the question content is positively related to the question response rate.

(c) *Polite Expression*: Balkan and Holmgren linked politeness strategies with perception, and found that teachers had positive opinions on students who politely emailed and expressed their willingness to cooperate with other students [19]. In social Q & A platform, polite expression can enhance users' platform identification, thereby promoting user participation [20]. According to the above research results, we believe that there is a certain relationship between polite expression and the higher question response rate. So we make a hypothesis:
H3: The use of polite expressions in question is positively related to the question response rate.

(d) *Expression of Urgency*: Many questioners in Zhihu have an urgent need for answers. The urgency of measuring a problem mainly starts from three aspects: emergency words [21], repeated punctuation, and repeated parenthesis [22].

In social Q & A platform, these expressions of urgency will have a certain degree of influence on the motivation of the respondent [14]. Among the data we collected, words similar to "emergency", "urgent", "wait", "online", etc., behavior like repeatedly asking several questions and repeated punctuation (???) appear frequently. So we make a hypothesis:

H4: The use of emergency expressions in question is positively related to the question response rate.

(e) *Affective Tendency Expression*: Emotion is a subjective consciousness of human beings, and it is the subjective reflection of the human brain on an objective existence [23]. Ekman used cross-cultural research to divide emotions into six basic types: joy, disgust, anger, surprise, fear, and sadness [24]. A question may be presented with a certain emotional tendency. We believe that positive and happy question content will enhance the users' response motivation while negative and sad question content will weaken the users' response motivation. So we make a hypothesis:

H5: Positive emotional tendencies in questions are positively related to the higher question response rate.

(f) *Posting Period*: According to survey data from iResearch.com, users of Zhihu have a good education background. 80.1% of them have a bachelor's degree or higher. In addition, most of the users' occupations are product manager, freelancer, programmer or lawyer. Based on the data above, we know that most of the users' working hours are from Monday to Friday, and their rest time is weekends. We suspect that the post time of the question is related to the response rate. During working hours from Monday to Friday, they are busy working and studying therefore rarely use Zhihu. So we make a hypothesis:

H6: The posting period of question is related to the response rate of the question. The response rate for questions raised during the weekdays will be lower.

4.2 What is the Characteristics of the Questioner?

(a) Reliability of The Questioner: Reliability of the questioner is critical to building trust and increasing influence in social Q&A platform. Previous researches mainly focused on the reliability evaluation of information, and the reliability of the information source will also affect the degree of acceptance of the information [14, 25]. Research shows that users who reshare information are more inclined to be elite users. Therefore, in Zhihu, the number of votes represent the reliability of the questioner. We suspect that the reliability of the questioner may affect the acceptance of the question. So we make a hypothesis:

H7: The reliability of questioner is positively related to the question response rate.

(b) Network Centrality of The Questioner: The centrality of the questioner's network includes centrality extroverted centrality and inward centrality [14]. Extroverted centrality refers to the number of other users that the questioner follows, and inward centrality refers to the number of people who follows the questioner. The higher extroverted centrality of the questioner, the more his response behavior and the higher inward centrality, the higher visibility of the question [26]. Paul (2011) concluded that the probability of receiving a reply is intrinsically related to the number of followers of the questioner. From a user's information in Zhihu, we can obtain the number of people he has followed and the number of his followers. We guess that the higher network centrality of the questioner, the higher probability that his question will be answered. We make a hypothesis:

H8: The extroverted centrality of questioner is positively related to the question response rate.

H9: The inward centrality of questioner is positively related to the question response rate.

(c) The Questioner's Degree of Social Learning: The concept of social learning is related to the rise of Internet Technology, Social Net-work, and Social Network Sites [27]. There are two main aspects of research on social learning: related research based on the perspective of knowledge networks and on social networks [28]. Social learning theory holds that observing the social learning process of other individuals can allow them to gain or lose a certain behavior [29]. We think that when a user who browses a lot of questions that have been asked and asks multiple questions by himself, he is repeatedly performing social learning. Through social learning, the users can slowly master the skills of asking questions, keep learning, and constantly change their questioning styles, making their questions tend to the higher response rate. So we make a hypothesis:

H10: The social learning level of questioner is positively related to the question response rate.

5 Research Method

5.1 Topic Selection

Through a survey by questionnaire, we randomly surveyed 568 users on topics of interest in Zhihu. A total of 524 valid samples were collected excluding samples with answer time less than 10 s and obvious filling errors. According to the survey results, we aggregated the percentage of people interested in each topic. We take topics in which the number of interested people accounts for more than 50% as research objects. The six specific topics are Education, Food, Photography, Law, Entrepreneurship and Health (Fig. 1).

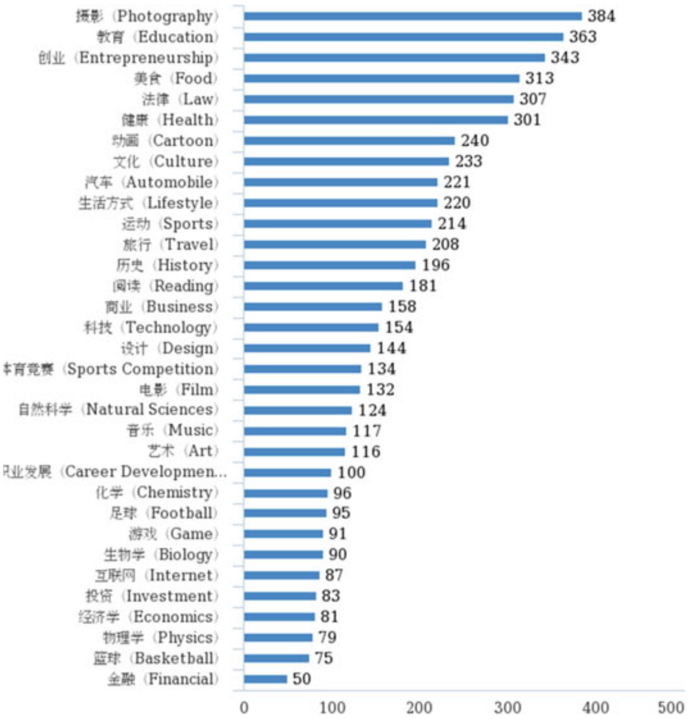


Fig. 1 Questionnaire about Zhihu topics of interest

5.2 Data Collection

By logging in to Zhihu's official website, and randomly selecting users who answer under a question, we crawled about 10,639 pieces of data, excluding 1477 pieces of incomplete information data. According to its topic classification, a total of 4293 pieces of data in 6 topics are selected.

5.3 Variable Measurement

See Table 1.

(a) *Carrier Richness*: According to the platform data, the carrier richness is divided into 0 and 1. If the question contains only plain text and "?", it is represented by 0. If there are other special symbols, pictures and links, etc., it is represented by 1.

Table 1 Variable symbol

		Variable name		Symbol
Independent Variables	Characteristics of the question	Carrier richness		CR
		Question length		QL
		Polite expression		PE
		Expression of urgency		EU
		Affective tendency expression		ATE
		Posting period		PP
	Characteristics of the questioner	Reliability of the questioner	The number of endorsements received	RQ
			The number of answers given by the questioner	
		Network Centrality of the questioner	Extroverted centrality	EC
			Inward centrality	IC
		The questioner's degree of social learning	The number of questions	QDSL
			The number of concerns	
Dependent variable		The situation of answer		Y
Constant		Constant		C

(b) *Question Length*: The downloaded data is stored with excel 2016. We use the LEN () function to find the character length (including punctuation) for each problem.

(c) *Polite Expression*: We process each sentence through text segmentation. If it contains "Thank you", "Thank you very much", "Please", "Excuse me", "It bothers you", "I'm sorry", "Trouble you", etc. 20 polite words and euphemisms expressing euphemisms, etc., we mark it as 1, otherwise we mark it as 0.

(d) *Expression of Urgency*: We process each sentence through text segmentation. If it contains "Emergency", "Urgent", "Wait", "Ask for help", "Online", etc. and repeated punctuation (for example, !!! ????), we mark it as 1, otherwise we mark it as 0.

(e) *Affective Tendency Expression*: This paper is based on the SnowNLP database and the Chinese word segmentation problem. We put the text data into a database, and the database contains the corresponding sentiment word weights. A weighted summation of these words yields emotional orientation results. The value ranges from 0–1. The closer to 1, the stronger the positive and happy emotional tendency. The closer to 0, the stronger the negative and sad emotional tendency. Around 0.5 is neutral.

(f) *Posting Period*: We use the MONTH () function to convert the collected time data into the day of the week. Mark Monday to Friday as 1 and weekends as 0.
(g) *Reliability of The Questioner*: In the user profile, there is a certification and achievements column that displays “Number of votes”. X = the number of votes received by the questioner ÷ the number of answers from the questioner. The X values of all sample data are sorted, and they are equally divided into 5 segments. Each segment takes the values at the two ends as intervals, and assigns 0–5 points to each interval. The score value represents the reliability of the questioner. The higher the score, the higher the reliability. When X = 0, the score is 0; when X is between 0 and 1, the score is 1; when X is between 1 and 3, the score is 2; when X is between 3 and 11, the score is 3; when X is between 11 and 80, the score is 4; when X > 80, the score is 5.
(h) *Network Centrality of The Questioner*: Extroverted centrality refers to the number of other users that the questioner follows, and inward centrality refers to the number of people who follows the questioner.
(i) *The Questioner’s Degree of Social Learning*: We think that when a user who browses a large number of questions and asks multiple questions by himself, he is repeatedly performing social learning. They slowly master the skills of asking questions, making their questions tend to have a high response rate.
(j) *Dependent variable measurement*: If the number of answers on topic A is 0, we mark Y = 0, otherwise it is 1.

5.4 Method of Prediction

A binary logistic regression model is used to verify the data of the influencing factors, and a predictive model of question response rate is constructed. The data set is divided into a test set and a training set, and a part of data is randomly selected as a test set to verify the accuracy of the prediction model. The remaining data is used as a training set to verify the influencing factors and build a prediction model.

6 Results and Analysis

6.1 Descriptive Results

We analyzed 6 topics and studied 4293 questions and found that the average overall number of answers was 3.642, and 31% of the questions asked were not answered; 55.7% of the respondents received less than 3 answers; It is found that the average character length for each question was 21.745. The average character length of questions that received a reply was 22.095, and the average character length of questions that did not receive a reply was 21.209. The average number of followers of each question was 21.829.

From the above data, we initially draw the following conclusions: Question response rate varied from topic to topic. There was the highest response rate in Education topic, but the lowest response rate in Law topic; The overall response was not very optimistic. Compared to the average number of followers, the average number of answers was less. Many people read questions rather than answered them; The average number of answers to each question was 3.642. Users' activity needs to be strengthened.

6.2 Models Building

We use IBM SPSS Statistics 24 software for binary logistic regression. All 10 variables are imported into the software, and the results are shown in Table 2. We test the effect of model fit. The −2 Log likelihood value is used to test the overall fit of the model. The value is 468.177, which is greater than the chi-square critical value of 18.307. So, it's reasonable to consider the model fits well. The closer the Nagelkerke goodness of fit values is to 1, the better the fit. We use the Omnibus test. The chi-square values of the steps, modules, and models are all greater than the critical value of 6, and the significance is far less than the critical value of 0.05. So the model coefficient test is passed. We give the test results of this model. The accuracy of the model is 83% for 100 data sets randomly selected. The results show that the model fits the whole well (Table 3). There is no significant difference between the model predictions and observations.

We test the hypothesis of the Health topic, and find that it violates hypothesis H5, H8, H9 and H10. The prediction process of the remaining 5 topics is the same as above. The models of the five topics can fit the whole well. There is no significant difference between the model predictions and observations. Based on the results, we

Table 2 Operation result

Variable	B	S.E	Wald	Df	Sig	Exp(B)
RQ	0.340	0.074	21.202	1	0.000	1.404
QDSL	0.000	0.000	0.139	1	0.709	1.000
CR	1.392	0.296	22.047	1	0.000	4.023
PE	1.784	0.262	46.217	1	0.000	5.954
EU	1.489	0.280	28.364	1	0.000	4.431
QL	−0.039	0.013	8.841	1	0.003	0.962
PP	−0.790	0.294	7.201	1	0.007	0.454
EC	0.000	0.000	0.368	1	0.544	1.000
IC	0.000	0.000	0.000	1	0.999	1.000
ATE	0.055	0.368	0.022	1	0.881	1.057
C	-0.338	0.468	0.521	1	0.471	0.713

have produced Table 4 to summarize the results of hypothesis verification in different topics.

6.3 *Model Prediction*

We set the prediction standard model as: $g(X) = \alpha_1 CR + \alpha_2 QL + \alpha_3 PE + \alpha_4 EU + \alpha_5 ATE + \alpha_6 PP + \alpha_7 RQ + \alpha_8 EC + \alpha_9 IC + \alpha_{10} QDSL + C$. According to the model parameters and the results of hypothesis testing, prediction models in different topics can be obtained (Table 5). We randomly select 100 sets of data under each topic to make predictions. We calculate the probability of observations and make predictions based on binary logistic regression. When the probability is less than 0.5, the prediction result is that the question is not answered; when the probability value is greater than 0.5, the prediction result is that the question at least gets 1 answer. The prediction accuracy rate in the 100 sets of prediction data was over 80%.

Table 3 Operation result

Omnibus inspection				Model summary				
	Chi-Square	Df	Sig	−2 Log Likelihood	Nagelkerke R square	Y	0	1
Step	190.213	10	0.000			g(X)	0	1
Block	190.213	10	0.000			Num	36	47
Model	190.213	10	0.000	468.177	0.431	Accuracy	83%	

Table 4 The hypothesis result of six topics

	Education	Food	Photography	Law	Entrepreneurship	Health
H1	√	√	√	√	√	√
H2	√	√	√	√	√	√
H3	√	√	√	√	√	√
H4	√	√	√	√	√	√
H5	×	√	×	×	×	×
H6	√	√	×	×	√	√
H7	√	√	√	√	√	√
H8	×	√	×	√	√	×
H9	√	×	√	√	×	×
H10	×	×	×	√	√	×

Table 5 The forecasting modle and modle accuracy

Topic	Forecasting mode	Accuracy (%)
Education	g(X) = 1.579CR − 0.037QL + 1.016PE + 0.927EU − 0.685PP + 0.237RQ + 0.0001IC + 0.418	86
Food	g(X) = 1.504CR − 0.057QL + 1.368PE + 0.532EU + 0.630ATE − 0.602PP + 0.567RQ − 0.232	87
Photography	g(X) = 1.645CR − 0.071QL + 1.154PE + 1.546EU + 0.487RQ + 0.0001EC − 0.736	82
Law	g(X) = 1.269CR − 0.039QL + 1.206PE + 1.874EU + 0.456RQ + 0.001EC + 0.0001IC + 0.0001QDSL − 1.394	90
Entrepreneurship	g(X) = 1.581CR − 0.043QL + 1.143PE + 1.679EU − 1.056PP + 0.438RQ + 0.001EC + 0.0001QDSL − 0.431	88
Health	g(X) = 1.387CR − 0.039QL + 1.789PE + 1.498EU − 0.770PP + 0.339RQ − 0.338	83

7 Conclusion

Reliability of the questioner will affect the question response rate of Zhihu. Usually highly reliable users will have a greater probability of receiving a response when they ask a question. if we want to get a greater probability of getting answers to questions in Zhihu, on the one hand, we should improve our reliability; on the other hand, we should add pictures and special symbols to the question to enhance the carrier richness of the question, minimize the length of the problem with clear content, pay attention to the politeness of the sentence when asking questions, and sometimes use some urgent words to get an answer quickly.

For Education topic, if more people have followed us, we may have a higher probability of receiving a response. We understand that the probability of receiving a response to a question posted at weekends is greater; For Food topic, in addition to choosing to post questions at weekends, positive emotional tendencies will raise the question response rate; For Photography topic, network centrality plays an important role.

Law is a hot topic in Zhihu, and the number of people we are followed by will affect the response rate of questions. For Entrepreneurship topic, the probability of receiving a response to a question posted at weekends is greater than a question posted on weekdays. Focusing on people who are interested in starting a business is very helpful to our response rate. Constantly learning and browsing other people's questions will also affect the question response rate; if we want to ask something about health, choosing to post questions at weekends will be a positive impact on improving the question response rate.

This paper analyzes the six topics response rates in Zhihu platform, hoping to provide some guidance and help for users (Tables 4 and 5).

When asking questions and increase the probability of a question being answered by changing several variables and paying attention to some details. However, there are two limitations in this article: (1) The variables selected for the prediction of the response rate are limited, and there may be more other variables that will affect the results; (2) The quality of Zhihu platform users is high. Therefore, the applicability to all platforms and topics needs further research.

Acknowledgements The research is supported by Science Foundation of China University of Petroleum, Beijing (No. 2462020YXZZ041).

References

1. Z. He, Z. Chen, S. Oh et al., Enriching consumer health vocabulary through mining a social Q&A site: a similarity-based approach. J. Biomed. Inform. **69**, 75–85 (2017)
2. Distribution of global social content sharing activities as of 2nd quarter 2016, by social network. Retrieved from. https://www.statista.com/statistics/283889/content-sharing-primary-social-networks-worldwide/, Accessed date: 14 March 2019
3. L. Wenyin, T. Hao, W. Chen et al., A web-based platform for user-interactive question-answering. World Wide Web **12**(2), 107–124 (2009)
4. S. Rafaeli, D.R. Raban, G. Ravid, How social motivation enhances economic activity and incentives in the Google answers knowledge sharing market. Int. J. Knowl. Learn. **3**(1), 1–11 (2007)
5. Y. Ding, Zhihu: the operation of the Q & A community [J]. Shanghai Informatization (09), 74–76 (2019)
6. Y. Shi, Analysis on the development status and improvement strategies of knowledge payment—Taking Zhihu as an example [J]. Commer. Econ. (05), 137–139 (2019)
7. J. Li, T. Zhang, Research on Influencing factors of users' perceived usefulness of knowledge sharing based on the social Q & A community—Taking Zhihu as an example [J]. Mod. Intell. **38**(04), 20–28 (2018)
8. L. Deng, Research on knowledge communication in the social Q & A community from the perspective of consumer culture [D]. Nanjing Normal University, (2017)
9. P. Liu, R. Lin, Research on knowledge sharing and communication behavior of "Zhihu" in the social Q & A community [J]. Libr. Inf. Knowl. (06), 109–119 (2015)
10. E. Choi, V. Kitzie, C. Shah, "10 Points for the best answer!" – Baiting for explicating knowledge contributions within online Q&A, in *Proceedings of the American Society for Information Science and Technology*. 50. https://doi.org/10.1002/meet.14505001101 (2013)
11. S. Oh, The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. J. Am. Soc. Inf. Sci. Technol. **63**, 543–557 (2012)
12. D. Dearman, K.N. Truong, Why users of Yahoo! answers do not answer questions [C].//*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 2010, pp. 329–332 (2010)
13. K.K. Nam, M.S. Ackerman, L.A. Adamic, Questions in knowledge in?: a study of naver's question answering community [C].//*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 2009, pp. 779–788 (2009)
14. D. Shengli, F. Shaoxiong, L. Jin, Prediction research on question response rate of social question answering platform—taking "Baidu Knows" as an example. Lib. Inf. Serv. **63**(10), 97–105 (2019)
15. E.A. Ramalho, J.J.S. Ramalho, Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models. Comput. Stat. Data Anal. **54**(4), 987–1001 (2010)

16. S.J. McMillan, Effects of structural and perceptual factors on attitude toward the website. J. Advertising Res. **43**(4), 400–421 (2004)
17. E.J. Brookes, The anatomy of a Facebook post. Study on post performance by Type, Day of Week, and Time of Day. [2018–10–01]
18. Applied Intelligence; Recent findings from Beijing Institute of Technology provides new insights into applied intelligence (User correlation model for question recommendation in community question answering). J. Robot. Mach. Learn. (2020)
19. S. Bolkan, J.L. Holmgren, "You are such a great teacher and I hate to bother you but…": instructors' perceptions of students and their use of email messages with varying politeness strategies. Commun. Educ. **61**(3), 253–270 (2012)
20. E. Joyce, R.E. Kraut, Predicting continued participation in ne- wsgroups. J. Comput.-Mediated Commun. **11**(3), 723–747 (2006)
21. E. Hellier, J. Edworthy, B. Weedon, K. Walters, A. Adams, The perceived urgency of speech warnings: semantics versus acoustics. Hum. Factors J. Hum. Factors Ergon. Soc. **44**(1), 1–17 (2002)
22. Y.M. Kalman, D. Gergle, CMC cues enrich lean online communication: the case of letter and punctuation mark repetitions (2010)
23. R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191
24. P. Ekman, T. Dalgleish, M. Power, *Handbook of Cognition and Emotion* (Wiley, Chichester, 1999)
25. C. Yaping, D. Xuebing, Effects of the content characteristics of virtual community consumer information on information sharing behavior. J. Inf. **33**(1), 200–206 (2014)
26. J. Weng, E.P. Lim, J. Jiang, et al., Twitter rank: finding topic-sensitive influential twitters, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (ACM, New York, 2010), pp. 261–270
27. R. Ferguson, S.B. Shum, Social learning analytics: five approaches, in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, ed. by S.B. Shum, D. Gasevic, R. Ferguson (ACM, New York , 2012), pp. 23–33
28. J. Duan, S. Yu, Construction of learning model based on social knowledge network. Res. Mod. Dist. Educ. **04**, 91–102 (2016)
29. C.H. Bates, An applied test of the social learning theory of deviance to college alcohol use in the context of a community enforcement campaign. Dissertations & Theses Gradworks (2013)

Content Recommendation of Tender Documents Based on Qualitative Characteristics

Tingting Zhou, Guiying Wei, and Ai Wang

Abstract Aiming at content recommendation of tender documents, this paper puts forward the case reuse and case modification algorithm of tender documents. First, according to the usage of clauses in tender cases, this paper uses non-interference sequence index to cluster similar tender cases and similar clauses, then based on which the reference samples and content modules of the tender documents were constructed. Finally, recommended value of reference samples and difference degrees between content modules were used respectively to realize content recommendation. This algorithm ensures the scientificity of the tender documents' preparation and the accuracy of the recommended content, and greatly improves the efficiency while reducing the scope of the recommended content.

Keywords Recommendation of tender documents · Case reuse · Case revision · Non-interference sequence index · Reference samples

1 Introduction

Tender work is a complex mental activity, during which a large number of text materials are produced. For example, the tender personnel shall write tender documents according to project characteristics, tender conditions, qualification requirements of the bidder, bid evaluation methods, technical standards and requirements, and the preparation of these documents also take a lot of time and energy [1, 2].

At present, there are not many literatures on the optimization methods for the preparation of tender documents. Ling emphasized the key points

T. Zhou · G. Wei (✉) · A. Wang
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: weigy@manage.ustb.edu.cn

T. Zhou
e-mail: 1105220939@qq.com

A. Wang
e-mail: wangai22222@126.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_22

and precautions of the preparation of the tender documents in her paper. In addition, the tender personnel need to learn relevant laws and regulations, technical and business knowledge widely. The content of these traditional tender documents compiled are reliable, but the time cost is high [3]. With the development of artificial intelligence technology in recent years, machine writing provides an effective way to replace artificial writing, Wang proposed an electronic system for preparation of tender documents, which divided the documents into three parts: text document, format document and attachment, improving the efficiency of the preparation of tender documents, And through self-defining the tender documents structure and template, the problem that the traditional preparation system can not be expanded in different industries is improved [4]. In order to distribute relevant tender information to suppliers selectively, Goswami uses naive Bayes classifier to perform 10 times cross validation experiments, and automatically classifies tender documents into several predefined technical categories [5]. Based on the extraction method, Tang proposed a way of machine writing for tender documents, which has a strong innovation in the application of natural language processing to the preparation of tender documents, However, this method is still in exploration, which is greatly disturbed by the unstable characteristics of the tender project itself, so it is generally applicable to standardized and modular text writing [6].

It can be seen from the previous theory that the traditional tender documents compilation system has been unable to meet the implementation of today's large-scale tender tasks, while these new methods have their own problems in scientificity and universality. Therefore, after analyzing the actual preparation process of tender documents, this paper establishes a content recommendation algorithm of tender documents based on non-interference sequence index clustering. It can provide recommended sequence of reference samples for tender documents preparation, or provide recommendation set of content modules further, which can complete the recommendation and revision of the solution for the new tender documents well. This method ensures the efficiency and scientificity of the preparation while optimizing the preparation process of tender documents.

2 The Non-interference Sequence Index

Definition 1 The non-interference sequence is a positive integer sequence, $M = (M_1, M_2, \ldots, M_k, \ldots)$, where, nth clause is greater than the sum of the previous (n − 1) clause, i.e., $M_n > \sum\limits_{k=1}^{n-1} M_k, n \geq 2$.

Definition 2 Suppose a data set C has m objects and n attributes, and these attributes are all binary attributes, which are recorded as $C_1, C_2, \ldots, C_k, \ldots, C_n$, then d_{ik} is the value of object $D_i (i \in [1, m])$ for attribute C_k. The non-interference sequence index of object D_i is: $q(D_i, M) = \sum_{k=1}^{n} d_{ik} * M_k$, where, M is a selected non-interference sequence.

Lemma 1 [7] For any two objects D_i and D_j in the binary attribute data set C, whose attribute values are $d_{i1}, d_{i2}, \ldots, d_{in}$ and $d_{j1}, d_{j2}, \ldots, d_{jn}\left(d_{ik}, d_{jk} \in \{0, 1\}, k \in [1, n]\right)$, respectively, which are recorded as $J_C(D_i) = (d_{i1}, d_{i2}, \ldots, d_{in}) = \{ J_{\mathrm{k}}(D_i)|k \in [1, n]\}$, $J_C\left(D_j\right) = \left(d_{j1}, d_{j2}, \ldots, d_{jn}\right) = \{ J_k(D_j)|k \in [1, n]\}$. The non-interference sequence index is $q(D_i, M) = \sum_{k=1}^{n} d_{ik} * M_k$ and $q\left(D_j, M\right) = \sum_{k=1}^{n} d_{jk} * M_k$. Thus if $q(D_i, M) = q\left(D_j, M\right)$, $J_C(D_i) = J_C\left(D_j\right)$, otherwise $J_C(D_i) \neq J_C\left(D_j\right)$.

Based on the above theories, the non-interference sequence index can be used to measure the difference between two objects. If the non-interference sequence index is the same, the two objects have strong similarity. In the next part of this paper, the contents of the tender documents will be clustered and recommended according to this method.

3 Reuse and Revision of Tender Documents

The reuse and revision of the tender documents are based on the completion of the tender case retrieval. At present, all the historical tender cases similar to the new problems in the case database have been collected. Now it is necessary to use the similar cases mentioned above to complete the recommendation of the solution for writing the tender documents, and the tender personnel can reuse directly or modify the historical solution [8].

3.1 The Concepts of Tender Documents' Recommendation

Definition 3 The document unit is the part of the document content that has logical independence and theme in the overall content of the tender documents.

Definition 4 The reference sample is a solution in which all the content in the tender case database is not the same, and denoting it as S_i.

Definition 5 Clause D_i is the minimum content unit with logical number in the tender document. The tender document is composed of several clauses naturally.

Definition 6 Content module D_p^g is a collection of clauses, i.e., $D_p^g = \{D_i | i \in [1, g]\}$.

Taking the document unit of bidder qualification requirements (DUBQR) as an example, Table 1 describes 3 tender cases and 3 solutions included in the cases. However, the solutions of C_1 and C_3 are identical, thus only 2 reference templates are included finally.

Definition 7 $J_k(D_i)$ is the judgment function of whether clause D_i is used. If clause D_i is used in tender case C_k, thus $J_k(D_i) = d_{ik} = 1$; otherwise, $J_k(D_i) = d_{ik} = 0$◦

Table 1 The sample: DUBQR' s solution

C_k	Content of clauses
C_1	(1) The bidder has independent legal personality and the ability to perform the contract (D_1) (2) Only manufacturer bids are accepted (D_2) (3) The bidder has ISO quality system certification (D_3) (4) The bidder has the financial, technical and production capacity or supply performance required for the performance of the contract, and meets the corresponding conditions specified in the tender document (D_4)
C_2	(1) The bidder has independent legal personality and the ability to perform the contract (D_1) (2) The manufacturer is expected to make direct investment in the project, but also accept the agent's bidding; if the agent bids, the manufacturer's sole authorization for the project shall be provided (D_5) (3) Different units in charge of the same person, or with a controlling or management relationship, may not participate in the same tender or the tender for the same tender project whose bid sections have not yet been divided (D_6) (4) The bidder has ISO quality system certification (D_3)
C_3	(1) The bidder has independent legal personality and the ability to perform the contract (D_1) (2) Only manufacturer bids are accepted (D_2) (3) The bidder has ISO quality system certification (D_3) (4) The bidder has the financial, technical and production capacity or supply performance required for the performance of the contract, and meets the corresponding conditions specified in the tender document (D_4)

According to the usage of the clauses, the usage of the content module D_i^g in the tender case C_k could be obtained, i.e. $J_k\left(D_i^g\right) = \{J_k\left(D_j\right) | j \in [1, g]\}$. Similarly, the solution (reference sample) can be expressed as $S_k = \{D_i | J_k(D_i) = 1\}$. It can be seen that defining the usage of clauses can transform the complicated natural language text (As shown in Table 1) into the binary matrix data structure (As shown in Table 2), that can be directly involved in the calculation. It is an important basis for the transformation from qualitative to quantitative.

As shown in Table 2, the connection between clauses can be established through their usage of the tender case. When the two clauses appear or disappear at the same

Table 2 Usage of clauses in DUBQR

	C_1	C_2	C_3
D_1	1	1	1
D_2	1	0	1
D_3	1	1	1
D_4	1	0	1
D_5	0	1	0
D_6	0	1	0

time in some tender cases, it indicates that the two clauses are related closely, and the usage of the two clauses is very similar in the actual preparation of tender documents, such as set $\{D_1, D_3\}$, $\{D_2, D_4\}$, $\{D_5, D_6\}$; On the contrary, such as the clauses D_2 and D_5, D_4 and D_6, they never appear at the same time, the similarity between them is low.

3.2 *Content Recommendation Mechanism*

Through interviewing experts and business personnel in the field of tender and analyzing the actual preparation process of tender documents, it is found that the core activity of the preparation of tender documents is to collect similar historical tender documents, then select the preparation templates of tender documents and revise them partially [9–11]. This paper intends to describe the compilation of factual document units objectively; And in terms of compiling empirical document units, we focus its recommended contents on the template (reference samples) and local content (content modules) of the tender document. Finally, we will study the construction of them and their recommended sequence.

Based on the above analysis, this paper proposes the content recommendation mechanism of tender documents (as shown in Fig. 1), which ensures the completeness and effectiveness of the tender experience recommendation.

3.3 *Reuse of Tender Documents*

3.3.1 Construct Reference Samples

Firstly, the basic concept of the construction of reference samples of tender documents is defined.

Definition 8 Tender case class CL_ρ^g is a collection of tender cases, i.e. $CL_\rho^g = \{C_k | k \in [1, g]\}$.

Theorem 1 In the tender case database, if and only if the non-interference sequence index of all tender cases in the tender case class is the same, there is only one reference sample in the tender case class. This is defined as the decision theorem of reference samples.

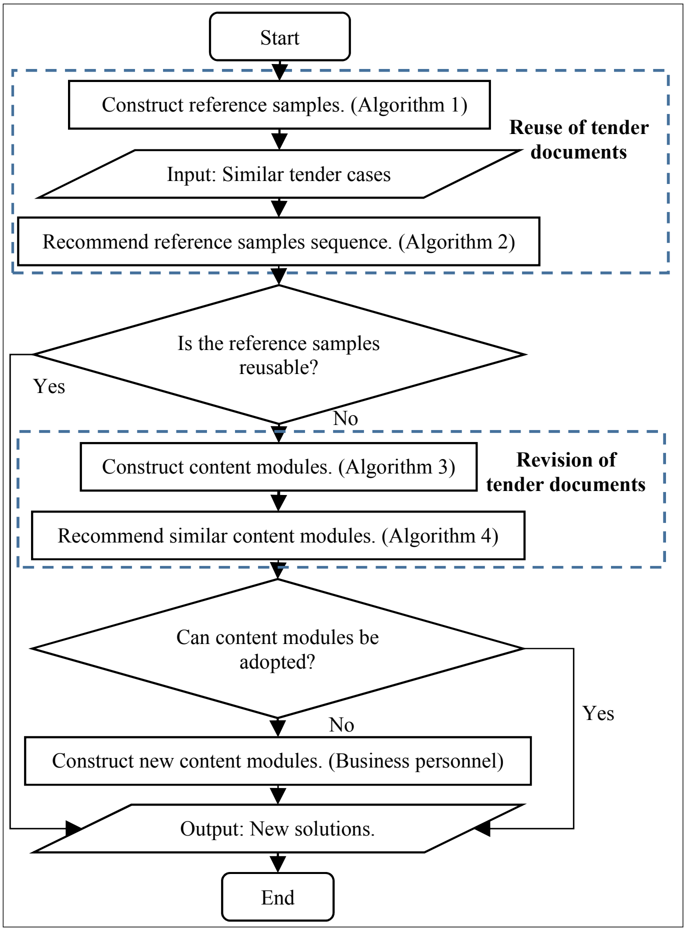

Start
Construct reference samples. (Algorithm 1)
Reuse of tender documents
Input: Similar tender cases
Recommend reference samples sequence. (Algorithm 2)
Is the reference samples reusable?
Yes
No
Construct content modules. (Algorithm 3)
Revision of tender documents
Recommend similar content modules. (Algorithm 4)
Can content modules be adopted?
Yes
No
Construct new content modules. (Business personnel)
Output: New solutions.
End

Fig. 1 The content recommendation mechanism of tender documents

The theorem shows that for any tender case class CL_{ρ}^{g}, and any non-interference sequence $M = (M_1, M_2, \ldots, M_i, \ldots)$ in the tender case database, assume any tender case $C_k, C_l \subset CL_{\rho}^{g}$ $(k, l \in [1, g]$ and $k \neq l)$, the solution of C_k is S_k, all clauses in the tender case database is $D = \{D_i | i \in [1, m]\}$, Thus if $q(C_k, M) = q(C_l, M)$, $S_k = S_l$, otherwise, $S_k \neq S_l$.

According to Lemma 1 and Theorem 1, the construction algorithm of reference samples of tender documents is proposed. The steps are as follows:

```
Algorithm 1 SamplesEstablish(C): The algorithm of constructing reference samples of tender documents
Input: usage of clauses $\{d_{ik}|i \in [1, m], k \in[1, n]\}$, arbitrary non-interference sequence $M$
Output: reference sample $S_p$, tender case class $CL_\rho^g$
Step 1: Calculate non-interference sequence index. Calculate the non-interference sequence index of each tender case $q(C_k, M)$
Step 2: Construct reference samples. Firstly, regard the solution $S_k$ of each tender case $C_k$ as a reference samples, meanwhile, regard each tender case $C_k$ as tender case class $CL_k^1$. Then, retrieve the sorting results of Step 1, merge the tender cases with the same index of noninterference sequence into a tender case class $CL_\rho^g$, and regard the solution of this class as a reference sample $S_p$
Step 3: Display result and output $S_p$, $CL_\rho^g$
```

3.3.2 Recommend Reference Samples Sequence

According to Fig. 1, the recommendation idea of reference samples of tender documents is proposed. First of all, all reference samples are identified from the tender cases which similar to the new problems, to ensure the relevance of the recommended content. Then, according to the frequency of the reference samples in the tender case set, the results of the descending order of the reference samples are recommended, which realizes the content recommendation of the reference samples in the tender documents and ensured the efficiency of recommendation work.

Definition 9 Frequency of reference samples $|S_i|_C$ is the frequency of reference samples S_i in the tender case set C.

Definition 10 $|C|$ is the total number of tender cases in tender case set C, the recommended value of reference samples in C is:

$$V_C(S_i) = |S_i|_C/|C| \tag{1}$$

Definition 11 The reference samples sequence SS is a sequence with strict order relations after sorting the reference samples set $S = \{S_i|i \in [1, M]\}$ according to some rules.

According to the recommendation idea of reference samples of tender documents, an algorithm of reference samples recommendation of tender documents is proposed. The steps are as follows:

Algorithm 2 *SamplesRecommend*(P): The algorithm of recommending reference sample sequence of tender documents

Input: the set of reference samples, $S_C = \{S_k | k \in [1, n]\}$ in tender case set C, similar tender case set C_{sim} of new problem P
Output: recommended sequence of reference sample S_{rec} for new problems P
Step 1: Filter reference samples. Identify the reference samples used by C_{sim} in S_C, it is expressed as $S_{Csim} = \{S_i | i \in [1, M], M \leq n\}$
Step 2: Assess reference sample value. Calculate the recommended value $V_{Csim}(S_i)$ of each reference sample in the S_{Csim}
Step 3: Sort reference samples. According to the recommended value of reference samples in **Step 2**, the reference sample set S_{Csim} is arranged in descending order, and the reference sample sequence $SS_{Csim} = (S_1, S_2, \ldots, S_M)$ is obtained
Step 4: Display result and output the recommended sequence of reference sample S_{rec} for new problem P

3.4 Revision of Tender Documents

3.4.1 Construct Content Modules

By analyzing the content recommendation mechanism in Fig. 1, it is found that the precondition for entering the recommendation process of content modules is that the business personnel have determined subjectively that all the recommended reference samples cannot be used directly, and need to modify the content of one of the reference samples. At this time, the tender case set similar to the new problem has been useless, so in order to ensure the accuracy of case revision, it is necessary to implement the recommendation of content modules from the whole tender case database.

Theorem 2 In the tender case database, if and only if the non-interference sequence index of all clauses in the content modules is the same, the content modules can divide all reference samples. This is defined as the decision theorem of content modules.

The theorem shows that, for any content module D_p^g and any non-interference sequence M in the tender case database $C = \{C_k | k \in [1, n]\}$, set the solution (reference sample) of C_k is S_k, assume any clauses $D_i, D_j \subset D_p^g$ ($i, j \in [1, g]$ and $i \neq j$), (1) If $q(D_i, M) = q\big(D_j, M\big)$ and $D_i \in S_k$, thus $D_p^g \in S_k$; (2) If $q(D_i, M) = q\big(D_j, M\big)$ and $D_i \notin S_k$, thus $D_p^g \notin S_k$; (3) If $q(D_i, M) > q\big(D_j, M\big)$ or $q(D_i, M) < q\big(D_j, M\big)$ and $D_i \in S_k$, thus $D_j \notin S_k$; (4) If $q(D_i, M) > q\big(D_j, M\big)$ or $q(D_i, M) < q\big(D_j, M\big)$ and $D_i \notin S_k$, thus $D_j \in S_k$.

According to lemma 1 and theorem 2, the construction algorithm of content modules of tender documents is proposed. The steps are as follows:

Algorithm 3 *ModulesEstablish*(C): The algorithm of constructing content modules of tender documents

```
Input: usage of clauses $\{d_{ik}|i \in [1, m], k \in [1, n]\}$, arbitrary non-interference sequence $M$
Output: content module $D_p^g$
Step 1: Calculate non-interference sequence index. Calculate the non-interference sequence index of each clause $q(D_i, M)$
Step 2: Construct content modules. Regard each clause $D_i$ as a content module $D_i^1$, then, merge the clause with the same index of noninterference sequence into a content module $D_\rho^g$
Step 3: Display result and output $D_p^g$
```

3.4.2 Recommend Similar Content Modules

Based on the clustering method, this part puts forward the idea of content module recommendation. The target module is the content module of the tender case database adopted to write the new solution. Based on the calculation of all content modules in the tender case database, content modules with the similar usage of the target module are recommended, which ensures the pertinence of the recommended content.

Definition 12 It is known that the usage of clauses D_i and D_j in the tender case database $C = \{\mathrm{C}_k|k \in [1, n]\}$ are represented as $J_C(D_i) = \{d_{ik}|k \in [1, n]\}$, $J_C(D_j) = \{d_{jk}|k \in [1, n]\}$, the difference degree between D_i and D_j is:

$$Sim_{doc}\left(D_i, D_j\right) = \sqrt{\sum_{k=1}^{n}\left|d_{ik} - d_{jk}\right|^2} \tag{2}$$

According to Theorem 2 and Algorithm 2, all clauses in the same content module are used the same in the tender case database, so a content module only needs to select one clause as a representative to calculate the degree of difference between the content modules.

Definition 13 As for content modules $D_p^g = \{D_i|D_i \in D_p^g, i \in [1, g]\}$ and $D_q^l = \left\{D_j|D_j \in D_q^l, j \in [1, l]\right\}$, the difference degree between D_p^g and D_q^l is:

$$Sim_{doc}\left(D_p^g, D_q^l\right) = Sim_{doc}\left(D_i, D_j\right) \tag{3}$$

Definition 14 The content modules whose difference degree meets a certain threshold are combined into a content module class $DM_x = \{D_p^g|D_p^g \in DM_x, x \in [1,\ \mathrm{R}]\}$, and content modules within the same category can be recommended to each other.

According to the recommendation idea of content modules of tender documents, an algorithm of recommendation of content modules of tender documents is proposed. The steps are as follows:

```
Algorithm 4 ModulesRecommend($D_p^g$) The algorithm of recommending similar content modules of tender documents
Input: content module set $D = \{D_p^g | p \in [1, P]\}$ in tender case set $C$, usage of content modules, $\{d_{pk} | p \in [1, P], k \in [1, n]\}$, recommended sequence of reference samples $S_{rec}$ for new problem P
Output: the recommended set of content modules $O(p)(p \in [1, P])$, the difference degree between content modules $OD_{sim}$
Step 1: Content modules divide the reference samples. Identify the recommended reference samples of new problems, and divide the content module set of each reference sample, $DC_x = \left(D_p^g, D_q^l, D_s^e, D_r^h\right)$
Step 2: Calculate the difference degree between content modules, $OD_{sim}$
Step 3: Construct content module class. Combine the content modules with the content module difference that satisfy $OD_{sim} < \beta$, and generate them into a content module class $DM_x, x \in [1, \text{R}]$
Step 4: Identify all target modules $D_p^g$ in Step 1, The content modules belonging to the same class as the target module are merged into its recommendated set $O(p)$
Step 5: The business personnel can select any of the recommended reference samples to add, delete or modify the content modules, and these operations can be performed through $O(p)$ of the target module to generate a new tender document
Step 6: Display result and output $O(p)$
```

4 Experiments and Results

Chossing the document unit of bidder qualification requirements (DUBQR) for example, this paper will verify the effectiveness content recommendation algorithm for the tender cases from the aspects of tender case reuse and tender case revison. 148 original tender documents of a certain company were selected as data sources for experiments, and 25% of the tender documents were randomly selected as the test set, and the remaining 75% were used as the training set.

4.1 Recommendation Process of Case Reuse

Input: new problem case 27, usage of clauses in tender documents $\{d_{ik} | i \in [1, 44], k \in [1, 111]\}$, the non-interference sequence$(1, 2, 4, 8, 16, \ldots, M_k, \sum_{k=1}^{n-1} M_k, \ldots, M_{44})$;

First of all, according to the usage of clauses and the non-interference sequence index of all tender cases in the training set, 27 tender case types and 27 corresponding reference samples are obtained.

Step1: Based on the case retrieval algorithm, the tender case set similar to the new problem is retrieved by using the case characteristics of the new problem, that is $\{C_{37}$, C_{53}, C_{70}, C_{72}, C_{73}, C_{82}, C_{86}, C_{93}, C_{99}, C_{100}, C_{101}, C_6, C_7, C_{11}, C_{12}, C_{13}, C_{18}, C_{23}, C_{25}, C_{26}, C_{27}, C_{29}, C_{30}, C_{32}, C_{33}, C_{59}, C_{95}, C_{103}, C_{108}, C_9, C_{10}, C_{21}, C_{22}, C_{49}, C_{105}, C_{87}, C_{54}, C_{65}, C_{104}, C_{109}, C_1, C_2, C_3, C_5, C_{14}, C_{17}, C_{20}, C_{24}, C_{34}, C_{35}, C_{38}, C_{39}, C_{41}, C_{50}, C_{51}, C_{62}, C_{66}, C_{67}, C_{68}, C_{69}, C_{71}, C_{74}, C_{75}, C_{77}, C_{78}, C_{80}, C_{81}, C_{83}, C_{85}, C_{89}, C_{92}, C_{94}, C_{102}, C_{110}, C_{40}, C_{42}, C_{96}, C_{15}, C_{16}, C_{45}, C_{46}, C_{47}, $C_{56}\}$, The non-interference sequence index is used to divide 10 tender case classes. Each case class has the same reference sample. Therefore, the reference sample set {1, 2, 3, 4, 5, 7, 12, 13, 23, 24} to be recommended is obtained;

Step2: Calculate the recommended value of each reference sample in the reference template set, that is (0.048, 0.012, 0.831, 0.024, 0.012, 0.024, 0.012, 0.012, 0.012, 0.012);

Step3: Use the reference value of the reference samples to sort in descending order, and intercept the first 3 reference samples as the recommended sequence(3, 1, 4), then, the business personnel can select the most ideal reference sample from the recommended sequence of reference samples to compile the tender documents.

Output: recommended sequence of reference samples for the new problem.

In the actual preparation of tender documents, as the volume of tender documents is too large, It is better to recommend a maximum of 3 pieces of content at a time, so the recommended sequence length of reference samples can be [1, 3] (Table 3).

4.2 Recommendation Process of Case Revision

4.2.1 Implement Algorithm 3

Input: usage of clauses in tender documents $\{d_{ik}|i \in [1, 44], k \in [1, 111]\}$, the non-interference sequence(1, 2, 4, 8, 16, …, M_k, $\sum_{k=1}^{n-1} M_k$, …, M_{44});

Calculate the non-interference sequence index $q(D_i, M) = \sum_{k=1}^{n} d_{ik} * M_k$ of each clause D_i, combine clauses of the same non-interference sequence index into a content module D_p^g;

Output: content modules, D_p^g (Results are displayed in Table 4).

4.2.2 Implement Algorithm 4

After case reuse, new problem cases without directly applicable reference samples need to enter the case revision to make some adjustments. As for the document unit of bidder qualification requirements, select the target module of new problem case 32 in the set for example. The operation steps are as follows:

Table 3 Recommended results of reference samples

New problem ID	Initial solution	S_{rec}	If recommend successfully? (1—success/0—fail)	New problem ID	Initial solution	S_{rec}	If recommend successfully? (1—success/0—fail)
1	3	(3, 1, 13)	1	20	13	(3, 1, 13)	1
2	3	(3, 1, 13)	1	21	3	(3, 1, 4)	1
3	3	(3, 1, 13)	1	22	3	(3, 1, 4)	1
4	3	(3, 1, 13)	1	23	5	(3, 1, 5)	1
5	3	(3, 1, 13)	1	24	4	(3, 1, 13)	0
6	1	(3, 1, 13)	1	25	17	(3, 1, 13)	0
7	1	(3, 1, 13)	1	26	3	(3, 1, 5)	1
8	1	(3, 1, 13)	1	27	3	(3, 1, 4)	1
9	1	(3, 1, 13)	1	28	1	(3, 1, 13)	1
10	3	(3, 1, 5)	1	29	19	(3, 1, 13)	0
11	8	(3, 1, 13)	0	30	3	(3, 1, 13)	1
12	3	(3, 1, 5)	1	31	3	(3, 1, 5)	1
13	3	(3, 1, 5)	1	32	2	(3, 1, 5)	0
14	9	(3, 1, 13)	0	33	3	(3, 1, 13)	1
15	10	(3, 1, 13)	0	34	25	(3, 1, 13)	0
16	5	(3, 1, 13)	0	35	1	(3, 1, 13)	1
17	15	(3, 1, 13)	0	36	26	(3, 1, 13)	0
18	3	(3, 1, 4)	1	37	3	(3, 1, 13)	1
19	5	(3, 1, 13)	0				

Table 4 Content module division of tender documents

Content module	Set of clauses	Content module	Set of clauses
D_1^1	1	D_{11}^2	13, 14
D_2^1	2	D_{12}^2	15, 16
D_3^1	3	D_3^1	17
D_4^1	4	D_{14}^1	18
D_5^1	5	D_{15}^1	19
D_6^1	6	D_{16}^2	22, 23
D_7^1	7	D_{17}^{10}	24, 25, 26, 27, 28, 29, 30, 31, 32, 33
D_8^4	8, 9, 10, 20	D_{18}^5	34, 35, 36, 37, 38
D_9^1	11	D_{19}^1	39
D_{10}^2	12, 21	D_{20}^5	40, 41, 42, 43, 44

Input: the recommended sequence of reference samples for new problem case 32, all set of content modules in the case database $D_p^g(p = [1, 21])$ and the usage of content modules;

Step1: The content module set of each reference sample in the new problem recommendation sequence of reference sample is divided, i.e. $DC_3 = (1, 2, 4, 5, 6, 7)$; $DC_1 = (1, 2, 4, 5, 6)$; $DC_{13} = (1, 2, 4, 5, 6, 9, 14)$;

Step2: Calculate the difference degree between content modules, OD_{sim}.

Step3: Through calculation, the difference degree between content modules belongs to [0, 10.39], so this paper intercepts the difference threshold $\beta = 10.39/2 = 5.19$, and selects the content modules that satisfy $OD_{sim} < 5.19$, and merge them into a content module class. Finally, 2 content module classes are obtained: $DM_1 = \{1, 2, 4, 5, 6, 7\}$, $DM_2 = \{3, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$;

Step4: Identify the target content module in Step1, combined with the classification of content module classes to obtain: $O(1) = \{2, 4, 5, 6, 7\}$; $O(2) = \{1, 4, 5, 6, 7\}$; $O(4) = \{1, 2, 5, 6, 7\}$; $O(5) = \{1, 2, 4, 6, 7\}$; $O(6) = \{1, 2, 4, 5, 7\}$; $O(7) = \{1, 2, 4, 5, 6\}$; $O(9) = \{3, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$; $O(14) = \{3, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20\}$;

Step5: The business personnel can select any of the recommended reference samples to add, delete, or modify the content modules. This operation can be performed by using the recommended set of the content modules $O(p)$ to generate a new tender document.

Output: the recommended set of the content modules $O(p)$.

Table 5 shows the results of the experiment on the remaining data after case reuse by using the case revision algorithm.

In case reuse, it can be seen from the indicator "If recommend successfully ?" (It indicates the recommendation is successful when there is an actual solution in the recommended sequence of reference samples.) that 67.57% of the tender cases can be successfully carried out through the case reuse recommendation of tender documents (as shown in Table 5); Then the unsuccessful cases are input into the case

Table 5 Recommended results of content modules for remaining test sets

New problem ID	Actual usage of content modules	Recommended reference samples and set of content modules	If recommend successfully? (1—success/0—fail)	Recommend location	Operation (Delete—D/Input—I/Alter—A)
11	new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	0		
14	1, 4, 6, 11, new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D/I
15	new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	0		
16	1, 2, 4, 5, 6, 9	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D
17	9, new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D/I
19	1, 2, 4, 5, 6, 9	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D
24	1, 3, 4, 5, 6, 7	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D/A
25	1, 4, 5, 6, 9, 14, new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D/I

(continued)

Table 5 (continued)

New problem ID	Actual usage of content modules	Recommended reference samples and set of content modules	If recommend successfully? (1—success/0—fail)	Recommend location	Operation (Delete—D/Input—I/Alter—A)
29	1, 4, 5, 6, 9, 14, new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	1	3	D/I
32	1, 4, 5, 6, 7	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 5 = {1, 2, 4, 5, 6, 9}	1	1	D
34	new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	0		
36	new	3 = {1, 2, 4, 5, 6, 7}, 1 = {1, 2, 4, 5, 6}, 13 = {1, 2, 4, 5, 6, 9, 14}	0		

revision model further, and compare the solution results with the actual results, It can be seen from the indicator of “If recommend successfully ?” (It indicates success when the actual usage of the content module can be obtained by inputting, deleting, and altering the recommended set of the target content module) that 21.62% of the tender cases can continue to modify the tender documents through the case revision model, and then obtain the tender documents successfully; So far, the overall rate of recommendation is 89.19%. While the randomly selected training set does not cover the whose usage of the clauses of the test set, thus four of the test data are not successful, the tender documents of these four cases must be prepared from scratch.

5 Conclusion and Prospect

In this paper, through combing the actual preparation process of the tender documents, the content recommendation mechanism of the tender documents is proposed; The non-interfering sequence index is used to construct recommended sequence of reference samples and recommended set of content modules of the tender documents, that can realize the process of preparing new tender documents by reusing directly or revising partially to get new tender documents. The results of real data show that the overall success rate of case reuse and case revision recommendation algorithm is as high as 89.19% for the content of tender documents with qualitative characteristics, and the recommendation effect is significant; at the same time, it helps business personnel reduce the scope of reference samples and content modules greatly, making the preparation process more efficient.

There are not many optimization methods for the preparation of tender documents in existing literature. This paper can enrich the recommendation methods of the content of tender documents to a certain extent. Next, the author of this paper will focus on the confirmation of the tender documents and introduce a new case learning process and evaluation system. By learning the solution and evaluation results of the new case into this algorithm, improve the accuracy and richness of the recommendation.

References

1. S. Wang, Factor analysis of the impact of the preparation of tender documents on the project cost. Architect. Eng. Technol. Des. **7**, 248+254 (2018)
2. H. Ying, Problems and Countermeasures in the compilation of project tender document. Build. Mater. Decoration **8** (2019)
3. J. Ling, Preparation of tender documents. Sci. Wealth **12**, 122 (2019)
4. R. Wang, Design and development of electronic tender documents compilation system—exploration of Baohua Tender’s Second Generation Electronic Tender Platform. China Tender **8**, 22–25 (2013)

5. S. Goswami, P. Bharwaj, S. Kapoor, Naïve Bayes classification of DRDO tender documents, in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, pp. 593–597 (2014)
6. J. Tang, Concepts and methods of machine writing tender documents. Tender Procurement Manage. **9**, 43–46 (2019)
7. S. Wu, Q. Wang, M. Jiang et al., Clustering algorithm of categorical data in consideration of sorting by weight. J. Univ. Sci. Technol. Beijing **35**(38), 1093–1098 (2013)
8. L. Xu, X. Li, M. Yu, Bayesian network modeling method based on case reasoning for emergency decision-making. J. Shanghai Normal Univ. Nat. Sci. **42**, 237–243 (2013)
9. S. Laryea, Quality of tender documents: case studies from the UK. Constr. Manag. Econ. **29**, 275–286 (2011)
10. J. Li, S.X. Pan, L. Huang, X. Zhu, A machine learning based method for customer behavior prediction. Tehnicki Vjesnik-Technical Gazette **26**(6), 1670–1676 (2019). https://doi.org/10.17559/TV-20190603165825
11. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik-Technical Gazette **25**(2), 528–535 (2018)

Environmental Uncertainty, Customer Concentration and Debt Financing Capacity

Jiaqi Ma and Lin Han

Abstract Customers are an important part of the supply chain, which have many aspects of the impact on the company's production and operation activities, but in China, few studies focus on corporate debt financing capacity involving the supply chain—customer relationship. According to the empirical analysis based on China's A-share manufacturing listed companies from 2014 to 2018, customer concentration can hugely improve corporate debt financing capacity, while this kind of significantly positive correlation exists only when environmental uncertainty is high. The results show that under the specific market conditions in China, large customers mean steady source of income and can provide strength certification for enterprises with their own reputation. From the bank's perspective, the positive effect on corporates' value outweighs the potential operating risks and damage to benefits, so understanding the supply chain can help banks make better credit decisions. However, different companies face different market situations, and customer concentration has different effects on debt financing capacity.

Keywords Environmental uncertainty · Supply chain risk management · Customer concentration · Debt financing capacity · Bank loan

1 Introduction

External financing is crucial to the development of the company. According to the theory of Pecking Order Theory, the cost of debt financing is lower and it is less restricted compared with equity financing. In addition, financial market is dominated by bank credit in China, and banks play an important role in the allocation of resources [1]. Information asymmetry is one of the fundamental problems of risk management of commercial loan. As an important part of the supply chain, customers are also an

J. Ma · L. Han (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: hanlin_0226@163.com

J. Ma
e-mail: 19120663@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_23

important economic resource for enterprises, which can compensate for information asymmetry to a certain extent and have important implications for debt financing capacity.

Using A-share manufacturing listed companies from 2014 to 2018 that have disclosed data of major customers as full sample, the paper tests the impact of customer concentration on corporate debt financing capacity. The results show that there is a significantly positive correlation between customer concentration and corporate debt financing capacity, however, this significantly positive correlation exists only when the environmental uncertainty is high. The result shows that in the specific market environment of China, the existence of large customers is considered to be able to promote the integration of the company's supply chain and represent the income guarantee of the company, thereby reducing the company's operating risk and transmitting positive signals to the banks, which can further improve the company's debt financing capacity. However, companies always operate in specific market environments, so the impact of customer concentration on corporate debt financing capacity changes in response to the changing environment.

The conclusions of this paper have both theoretical and practical significance. Firstly, it explores the relationship between the supply chain-customers and banks, and provides certain theoretical basis for the development of supply chain finance. Secondly, it explains the importance of large customers to corporate's debt financing, and provides solutions for companies with poor financing capacity, which can strengthen the management of the supply chain, attract high-quality customers, and establish good and stable customer relationships.

2 Literature Review and Research Hypothesis

2.1 Customer Concentration and Debt Financing Capacity

Customer concentration can affect the operation of the enterprise, and has an important impact on the enterprise, which has both positive and negative sides.

On the one hand, customer concentration increases corporate risk and adversely affects business performance. Enterprises with concentrated customers may face higher operating risk, because if major customers are in financial distress, facing bankruptcy, or changing suppliers, then companies that depend heavily on large customers will lose large orders, resulting in a sharp decline in sales revenue. Customer concentration can also do damage to business performance because they spend too much resources maintaining relationships with large customers, for example, the bargaining power of major customers increases with the customer concentration [2].

On the other hand, the contemporary market competition has changed from competitions among enterprises to competitions among supply chains. Bankers consider the corporate supply chain as an important factor in the credit decision-making process [3]. The existence of large customers in China is considered to be beneficial to the integration of the supply chain, and be able to transmit positive signals to the market [4]. The supply chain profoundly affects the economic benefits of each company among it. The stable relationship with major customers is the guarantee of a company's main source of income and its performance.

Taken together, I make the following assumptions:

Hypothesis 1a: The higher the customer concentration, the stronger the company's bank loan capacity.

Hypothesis 1b: The higher the customer concentration, the weaker the company's bank loan capacity.

2.2 Environmental Uncertainty, Customer Concentration and Debt Financing Capacity

A high degree of environmental uncertainty can exacerbate information asymmetry [5]. In addition, higher environmental uncertainty transmits negative signals to the capital market, leading to higher corporate financing costs [6]. At this time, customer concentration, an indicator that can weaken information asymmetry, will send signals to banks. In most cases, large customers are the guarantee of sales revenue, especially when the environment is dynamic, it is important for a company to be able to maintain stable profitability. Therefore, large customers can send positive signals to banks, making banks more inclined to companies with high customer concentration, thus improving debt financing capacity. This hypothesis can be stated as follows:

Hypothesis 2: When the degree of environmental uncertainty is high, the role of customer concentration in promoting corporate debt financing capacity is more significant.

3 Research Design

3.1 Empirical Model

To evaluate the impact of customer concentration on corporate debt financing capacity and explore whether environmental uncertainty affects the relationship between customer concentration and corporate debt financing capacity, I specify the regression models. The regression analysis of model (1) using full sample tests hypothesis 1. The regression analysis of model (2) under full sample and the grouped regression of model (1) according to the level of environmental uncertainty are used to test hypothesis 2. In addition, industry and year are controlled in the model.

The models can be written as follows:

$$\begin{aligned} LoanT_{it} = {} & \alpha_0 + \alpha_1 CC_{it} + \alpha_2 ROA_{it} + \alpha_3 Size_{it} + \alpha_4 LEV_{it} \\ & + \alpha_5 Growth_{it} + \alpha_6 OCF_{it} + \alpha_7 Top1_{it} + \alpha_8 PPE_{it} \\ & + \alpha_9 Outdir_{it} + Industry + Year + \varepsilon_{it} \end{aligned} \tag{1}$$

$$\begin{aligned} LoanT_{it} = {} & \alpha_0 + \alpha_1 CC_{it} + \alpha_2 HEU_{it} + \alpha_3 CC_{it} \\ & \times HEU_{it} + \alpha_4 ROA_{it} + \alpha_5 Size_{it} + \alpha_6 LEV_{it} + \alpha_7 Growth_{it} \\ & + \alpha_8 OCF_{it} + \alpha_9 Top1_{it} + \alpha_{10} PPE_{it} + \alpha_{11} Outdir_{it} \\ & + Industry + Year + \varepsilon_{it} \end{aligned} \tag{2}$$

The dependent variable of model (1) is LoanT, which is used to measure the size of bank loan. CC is the explanatory variable, which is used to measure the customer concentration. The remaining variables are control variables, and ε is the random error term. When the coefficient α1 of CC is positive and significant, it is assumed that H1a is established; otherwise, it is assumed that H1b is established.

Model (2) is based on model (1) adding HEU and the interactive variable CC × HEU, where HEU is a moderator and is used to explore whether environmental uncertainty affects the relationship between customer concentration and corporate debt financing capacity. The remaining variables are consistent with model (1). If the coefficient of the interactive variable CC × HEU is significant and positive, it indicates that the environmental uncertainty is the moderating variable of the two.

3.2 Variable Definitions

(1) *Bank Loan Size*
The dependent variable is the bank loan size (LoanT), referring to existing Chinese literature [7], LoanT is the ratio of the sum of short-term and long-term borrowing balances to the total assets at the end of the period.

(2) *Customer Concentration*
The explanatory variable is customer concentration (CC). Scholars in China generally measure customer concentration differently from foreign countries, because domestic and foreign countries have different requirements for accounting information disclosure [8]. The method generally adopted in foreign countries is to treat the enterprises with sales higher than 10% of the total sales as the main customers, and then use the sum of the sales ratio of the main customers to measure the customer concentration. Since 2007, China Securities Regulatory Commission has required listed companies to disclose the proportion of the top five customers in the company's total sales in their annual reports. Most Chinese scholars use the sum of the top five customers' operating income ratio as the measure of customer concentration, which is just the paper uses [9].

(3) *Environmental Uncertainty*
The environmental uncertainty (EU) is measured by excluding the coefficient of variation of the normal growth part of sales revenue in the past five years and adjusted by the industry, and then the EU is calculated by year. At last, a dummy variable of environmental uncertainty HEU is set to 0 if EU is less than the median, otherwise, it is set to 1 [6].

(4) *Control Variables*
With reference to the existing literature on bank loan capacity, the following control variables are set:

   (a) *Enterprise Size (Size)*
   It is expressed by the natural log of the total assets at the end of the period. In general, the scale of an enterprise represents its strength. Without a good management system, it is difficult for a company to survive, let alone expand its scale. Therefore, the larger the scale of a company, the better the management system and strength of the company.

   (b) *Return On Assets (ROA)*
   It is expressed as the ratio of net profit for the year to total assets at the end of the period. This indicator effectively reflects the profitability of the enterprise, and combines two important indicators in the income statement and balance sheet, which has an important impact on the company's access to bank loan.

   (c) *Asset-liability Ratio (LEV)*
   It is expressed as the ratio of total debt to total assets. When financial leverage is maintained at an appropriate level, it can play a positive role in expanding the scale of corporate income. Conversely, it can increase the financial risk of the enterprise, and it is also one of the indicators that banks review before making credit decisions.

   (d) *Corporate Growth (Growth)*
   It is divided by the difference between the operating income of the current year and the previous year divided by the operating income of the previous year. The faster the company's revenue grows, the better the company's future development. On the other hand, the growth of the company requires to raise funds. It helps banks understand the real need of funds to the company.

   (e) *Operating Net Cash Flow (OCF)*
   It is expressed as the ratio of the net cash flow from operating activities to the total assets at the end of the period. The more the company's operating net cash flow, the less demand for bank loan.

   (f) *Equity Concentration (Top1)*
   It is expressed by the shareholding ratio of the largest shareholder. The higher the concentration of equity, the higher the cost of corporate borrowing [9]. This means that concentration of equity reduces the company's ability to borrow money from banks.

(g) *Fixed Asset Ratio (PPE)*
It is expressed as the ratio of the net value of fixed assets to the total assets at the end of the period. When the bank conducts credit risk assessment, in order to reduce the risk that the company cannot repay at maturity, it will examine whether the company has sufficient assets to be pledged. Taking account into the bank's emphasis on the company's asset structure, it may affect the bank's ability to raise funds.

(h) *Proportion of Independent Directors (Outdir)*
It is expressed as the proportion of independent directors of the company to all directors. Corporate governance characteristics, one of which is exactly the ratio of independent directors, affect the cost of bank borrowing, and accordingly affect the ability of enterprises to obtain bank borrowings [10].

3.3 Data Source and Sample Processing

The paper uses a sample of A-share listed companies in Shenzhen Stock Exchange and Shanghai Stock Exchange from 2014 to 2018. After screening, the final total sample is data of 826 companies and annual data of 4,130 companies.

I use the following principles to screen the initial sample:

- Screen out ST companies to ensure the validity of the data.
- Exclude financial companies from the sample due to the special nature of the financial industry.
- Exclude listed companies that lack data or exist abnormal data on research variables to ensure the rationality of research results.
- Exclude companies not listed before 2010, because the variable of environmental uncertainty requires the company's income data for the past five years. The research year of this paper is 2014–2018, so it should be listed at least in 2009.

In addition, in order to rule out the effects of extreme values, tailing processing is performed on all continuous variables in this paper at the level of 1 percentile before data processing.

4 Empirical Analysis

4.1 Descriptive Statistics

It can be seen from Table 1 that the bank loan of the sample enterprises in the study accounted for an average of 18% of the total assets, with a median of 16%, which indicates that some of the sample enterprises have slightly more loans. In terms of customer concentration, the average sales ratio of the top five customers of

Table 1 Descriptive statistics of the main variables

Variables	Observations	Mean	Std. deviation	Min	Median	Max
LoanT	4130	0.180	0.140	0.000	0.160	0.850
CC	4130	0.290	0.220	0.020	0.220	1.000
HEU	4130	0.470	0.500	0.000	0.000	1.000
Size	4130	22.620	1.330	19.160	22.460	27.780
ROA	4130	0.030	0.060	−0.720	0.030	0.590
LEV	4130	0.500	0.200	0.030	0.510	0.990
Growth	4130	0.140	0.370	−0.910	0.080	2.890
OCF	4130	0.040	0.070	−0.500	0.040	0.660
Top1	4130	0.340	0.150	0.030	0.320	0.890
PPE	4130	0.240	0.180	0.000	0.210	0.950
Outdir	4130	0.370	0.060	0.180	0.360	0.710

China's listed companies is 29%, with a minimum value of 0.02% and a maximum value of 100%. It can be seen that customer characteristics of companies have large differences. But overall, the customer concentration of the sample companies is moderate. The mean and median values of the remaining variables are equal or less different, and there are no outliers in the maximum and minimum values.

4.2 Regression Results

(1) *Customer Concentration and Debt Financing Capacity*

The paper explores the impact of customer concentration on debt financing capacity through the OLS multiple regression method. The first column of Table 2 shows the regression results of model (1) testing hypothesis 1. It can be seen that the correlation coefficient of customer concentration (CC) is 0.043, and it is significant at the level of 1%, indicating that there is a significantly positive correlation between bank loan size and customer concentration, that is, the higher the customer concentration, the stronger the company's ability to borrow money from bank. The empirical analysis results are consistent with Hypothesis 1a, that is, Hypothesis 1a is verified.

(2) *Environmental Uncertainty, Customer Concentration and Debt Financing Capacity*

Two steps are used to test Hypothesis 2, which is to test whether there is a difference in the significance of the positive correlation between customer concentration and bank borrowing financing capacity under different levels of environmental uncertainty. The results are shown in Table 2. First, a multiple regression analysis is performed on the entire sample according to model (2). It can be seen from the second column of Table 2 that the coefficient of the interactive variable CC × HEU is 0.0224,

Table 2 Environmental uncertainty, customer concentration and debt financing capacity

Variable	Customer concentration and debt financing capacity	Environmental uncertainty, customer concentration and debt financing capacity		
	Full sample	Full sample	HEU = 1	HEU = 0
	(1)	(2)	(3)	(4)
CC	0.0430***	0.0290**	0.0557***	0.0067
	(0.0110)	(0.0130)	(0.0177)	(0.0174)
HEU		−0.0045		
		(0.0040)		
CC × HEU		0.0224**		
		(0.0111)		
Size	0.0156***	0.0153***	0.0036	0.0139**
	(0.0036)	(0.0036)	(0.0062)	(0.0063)
ROA	−0.0774***	−0.0775***	−0.1080***	−0.0513
	(0.0217)	(0.0217)	(0.0309)	(0.0399)
LEV	0.4630***	0.4630***	0.4460***	0.5130***
	(0.0120)	(0.0120)	(0.0190)	(0.0197)
Growth	−0.0076***	−0.0078***	−0.0062*	−0.0037
	(0.0028)	(0.0028)	(0.0035)	(0.0091)
OCF	−0.1540***	−0.1540***	−0.1520***	−0.1720***
	(0.0162)	(0.0162)	(0.0247)	(0.0263)
Top1	0.0259	0.0250	0.0318	-0.0377
	(0.0193)	(0.0193)	(0.0338)	(0.0307)
PPE	0.0627***	0.0626***	0.0476*	-0.0222
	(0.0160)	(0.0160)	(0.0282)	(0.0236)
Outdir	−0.0448	−0.0433	−0.1080**	0.0136
	(0.0284)	(0.0284)	(0.0513)	(0.0360)
Industry	Control	Control	Control	Control
Year	Control	Control	Control	Control
Constant	−0.3830***	−0.3710***	−0.1120	−0.5230***
	(0.0881)	(0.0885)	(0.1510)	(0.1570)
Observations	4,130	4,130	1,927	2,203
Ajust_R2	0.2640	0.2650	0.1190	0.1450

*, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively, in a two-tailed test

and it is significant at the 5% level, that is, the environmental uncertainty has a significant impact on the relationship between customer concentration and bank loan size. In other words, the higher the environmental uncertainty, the more significant the relationship between customer concentration and debt financing capacity. That is to say, although environmental uncertainty increases the dependence of enterprises on large customers and corporate risks to a certain extent, banks tend to show more trust to companies with high customer concentration with the increase of environmental uncertainty.

Then, the full sample data are divided into two subsamples based on the level of environmental uncertainty. The model (1) is used to perform a group regression analysis according to two groups of data. The results are shown in the third and fourth columns of Table 2. It can be seen that when HEU = 1, that is, when the enterprise is of high level of environmental uncertainty, the coefficient of customer concentration is 0.0557 and is significant at the level of 1%. At this time, customer concentration and debt financing capacity show a significantly positive correlation. However, when HEU = 0, that is, when the enterprise is at low level of environmental uncertainty, the coefficient of customer concentration is 0.00665, and not statistically significant. Therefore, only in the high-level uncertainty group, there is a significantly positive correlation between the two.

4.3 Robustness Test

The robustness test is performed by using the lagging period of the dependent variable. The results are the same as the regression results above.

Considering that the company's current bank loan balance does not fully reflect the borrowing capacity of the current year, because these loans are not generated in a single year. In addition, corporate governance factors on bank loans have lagged effects [9]. Therefore, in order to avoid the endogenous problems caused by the lagging impact of customer concentration on bank loan, I attempt to measure the dependent variable with one lag of the bank loan size (LoanT_lag). The results are shown in Table 3.

The results of the robustness test for Hypothesis 1 are shown in the first column of Table 3. The estimated coefficient of customer concentration is statistically significant at a level of 5%, with a coefficient of 0.0340. That is, the higher the customer concentration, the larger lagged bank loan size. Despite the decrease in significance, the results again validate Hypothesis 1a.

When conducting a robustness test on the relationship among environmental uncertainty, customer concentration, and debt financing capacity, I still use the same methods above. The results of the robustness test are shown in the last three columns of Table 3.

Firstly, regression analysis is performed on the entire sample. The coefficient of the interactive variable CC × HEU is 0.032, and it is statistically significant at the level of 5%, indicating that the regulatory effect of environmental uncertainty is

Table 3 Robustness test

Variable	Customer concentration and debt financing capacity	Environmental uncertainty, customer concentration and debt financing capacity		
	Full sample	Full sample	HEU = 1	HEU = 0
	(1)	(2)	(3)	(4)
CC	0.0340**	0.0128	0.0585**	0.0093
	(0.0155)	(0.0183)	(0.0248)	(0.0287)
HEU		-0.0076		
		(0.0054)		
CC × HEU		0.0320**		
		(0.0148)		
Size	0.0315***	0.0313***	0.0300***	0.0343***
	(0.0053)	(0.0053)	(0.0093)	(0.0100)
ROA	−0.0022	−0.0035	−0.0332	−0.1070
	(0.0299)	(0.0299)	(0.0435)	(0.0659)
LEV	0.1160***	0.1160***	0.0897***	0.1340***
	(0.0166)	(0.0166)	(0.0263)	(0.0280)
Growth	0.0006	0.0003	0.0008	−0.0135
	(0.0039)	(0.0039)	(0.0049)	(0.0141)
OCF	−0.1440***	−0.1440***	−0.1590***	−0.1030***
	(0.0218)	(0.0218)	(0.0333)	(0.0371)
Top1	−0.0159	−0.0167	−0.0173	−0.0853*
	(0.0264)	(0.0264)	(0.0491)	(0.0441)
PPE	−0.0068	−0.0073	−0.0398	0.0149
	(0.0220)	(0.0220)	(0.0394)	(0.0355)
Outdir	−0.0263	−0.0246	−0.0756	0.0389
	(0.0381)	(0.0380)	(0.0695)	(0.0500)
Industry	Control	Control	Control	Control
Year	Control	Control	Control	Control
Constant	−0.5130***	−0.5130***	−0.4200*	−0.7960***
	(0.1280)	(0.1280)	(0.2520)	(0.2460)
Observations	3304	3,304	1,547	1,757
Ajust_R2	0.0974	0.0990	0.1110	0.0910

*, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively, in a two-tailed test

significant. Then the group regression is based at high and low level of environmental uncertainty. The results obtained are consistent with the regression results above, that is, only in the high level of environmental uncertainty group, the coefficient of customer concentration and debt financing capacity is significantly positive.

5 Conclusion

Using the sample of all the A-share listed companies from 2014 to 2018 in China, the paper examines the impact of customer concentration on corporate debt financing capacity, and influence of environmental uncertainty on the relationship between customer concentration and corporate debt financing capacity. Overall, our results suggest that in the specific market environment of China, the existence of large customers is considered by the banks to be able to facilitate the integration of enterprise supply chain and provide income guarantees, so it can transmit positive signals to the banks, thereby reducing the bank's credit risk and improving corporate debt financing capacity. Moreover, our findings imply that the positive effect of large customers only occurs when the environmental uncertainty the company facing is high. The possible reasons could be that the company's earnings are more volatile, and the operating risk increases at that time. As a consequence, the credit decisions made by banks tend be more cautious. However, major customers are the guarantee of income, so a company with a high degree of customer concentration will generally be considered to have stable profitability, which can pass positive signals to banks, thereby improving the company's debt financing capacity. In the contrary, when the level of environmental uncertainty is low, the company's operating risk is also at a lower level, so the company's dependence on large customers is relatively smaller. Considering from the bank's perspective, other financial indicators of the company may be more important.

Acknowledgements This article is supported by "the Fundamental Research Funds for the Central Universities (2019YJS058)". Thanks for the support.

References

1. X.X. Liu, X.Y. Zhou, Examination of the allocation relationship between financial resources and the real economy—also on the causes of economic structure imbalance. Fin. Res. **02**, 57–70 (2011)
2. Y.J. Tang, Suppliers, distributors' bargaining power and company performance: empirical evidence from listed Chinese manufacturing companies in 2005–2007. China Ind. Econ. **10**, 67–76 (2009)
3. D. Wang, Z. J. Liu, and Z. P. Zhao, Supply chain relations and bank loans—analysis based on supplier/ customer concentration. Account. Res. **10**, 42–49 + 96 (2016)

4. X.Y. Wang, P. Wang, J.P. Zhang, Customer concentration and audit costs: customer risk or supply chain integration. Audit Res. **06**, 72–82 (2014)
5. D. Ghosh, L. Olsen, Environmental uncertainty and managers' use of discretionary accruals. Acc. Organ. Soc. **02**, 188–205 (2009)
6. H.H. Shen, The impact of environmental uncertainty on earnings management. Audit Res. **01**, 89–96 (2010)
7. J.G. Zhu, F.C. Han, Z.F. Lu, Industrial policy, bank connectivity and corporate debt financing: an empirical study based on A-share listed companies. Fin. Res. **03**, 176–191 (2015)
8. T. Li, M. Li, Factors influencing the foreign debt safety: an empirical study. Econ. Comput. Econ. Cybern. Stud. Res. **53**(4), 259–274 (2019)
9. Z.L. Chen, Customer concentration, government intervention and corporate risk. Account. Res. **11**, 23–29 + 95 (2016)
10. L.J. Yao, M. Luo, D.L. Xia, Corporate governance and bank loan financing. Account. Res. **08**, 55–61 + 96 (2010)

OD-HyperNet: A Data-Driven Hyper-Network Model for Origin-Destination Matrices Completion Using Partially Observed Data

Yuxuan Xiu, Wanda Li, Jing Yang (Sunny) Xi, and Wai Kin (Victor) Chan

Abstract Estimating the inter-city population flow is critical for modeling the spread of COVID-19. However, for most cities, it is difficult to extract accurate population numbers for inflow and outflow. On the other hand, mobile carriers and Internet companies can estimate the distribution of population flow by tracking their users; but their data only cover part of the travelers. In this paper, we present a data-driven hyper-network model to aggregate these two types of data and complete the inter-city OD matrix. We first propose a cross-layer breadth-first traversal algorithm to estimate the inflow and outflow population of each city, then complete the OD matrix with an optimization model. Our experiments on a real-world dataset prove the accuracy and efficiency of our model.

Keywords Hyper-network · Data-driven modeling · Origin-destination matrix

1 Introduction

The current COVID-19 outbreak is one of the greatest challenges that humanity has ever faced, and consequently, many new research questions have surfaced. Population movement between regions, on which many epidemic models rely, is an important factor for modeling the spread of COVID-19 [1, 2]. For the first wave of the COVID-19 epidemic that started in Wuhan, China, it has been found that the geographic distribution of COVID-19 infections can be rather accurately predicted by the population outflow from Wuhan [3]. For the new clusters of COVID-19 cases emerging

Y. Xiu · W. Li · J. Y. Xi · W. K. Chan (✉)
Shenzhen Environmental Science and New Energy Technology Engineering Laboratory,
Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,
Tsinghua University Shenzhen, Shenzhen, People's Republic of China
e-mail: chanw@sz.tsinghua.edu.cn

Y. Xiu
e-mail: yuxuanxiu@gmail.com

W. Li
e-mail: wdli10@outlook.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_24

in Jilin and Beijing recently, population movement data are also fundamental in estimating the risk of the epidemic spreading to other cities.

The number of people traveling between different places can be specified using the Origin Destination (OD) matrix, where the element n_{ij} represents the number of people moving from region i to region j. As an important method in the field of transportation and logistics, OD matrix estimation has long been studied. A characteristic of the OD matrix estimation problem is that it relies heavily on the available data. Researchers have proposed a variety of OD matrix estimation algorithms to handle different types of input data. Before pervasive computing devices (e.g. mobile phone and GPS devices) are widely used, the most commonly used data source is traffic counts. For this type of input data, only the traffic flows at specific positions are observed, while the trajectories of travelers are missing. A commonly used approach is to impose some assumptions (e.g., the route choice pattern) to the OD matrix. Transport demand models, such as the gravity model [4] and the opportunity model [5], are calibrated based on the traffic counts to perform the estimation. With the development of mobile devices, digital footprints are available nowadays, enabling data-driven OD estimation algorithms. Data-driven algorithms usually leverage cell phone data [6] or GPS data [7] to estimate the OD matrix, focusing on tracking travelers based on the raw positioning data.

This paper aims to tackle a different OD matrix completion problem, which is encountered in modeling the inter-city spread of COVID-19 in China. On one hand, only a small number of cities can provide accurate traffic counts (i.e., the total number of inflow and outflow population), as most local governments are unable to track all of the origins or destinations of the population flow. On the other hand, mobile carriers (e.g., China Mobile) and Internet companies (e.g., Baidu) can track their users based on the positioning data, thus calculating the distribution of population inflow and outflow of each city. However, they may fail to obtain the total number of people moving in and out of cities, since the positioning data can only cover their own users.

In this paper, we assume that the following two types of data are available: (1) the accurate total population outflow (or inflow) of one or a few cities, and (2) the distribution of population inflow and outflow of each city. This paper aggregates such information to complete the inter-city OD matrices, which is rather different from the two typical OD matrix estimation problems mentioned earlier. On one hand, extracting trajectories of travelers is beyond the scope of this paper. Rather, we assume that such data is already provided by mobile carriers or Internet companies. On the other hand, traffic counts-based models are not very suitable for our problem, where the aggregation of existing information alone is sufficient to complete the OD matrix. Additional inferences and assumptions are not necessary.

One powerful model to aggregate information from multiple data sources is the hyper-network model [8–11]. Its hyper-network structure could transform and combine complicated mathematical relationships between data into one structure. A standard hyper-network model is composed of several single layers and inter-layer connections between them. Data from the same information source form a single-layer network, where each data element serves as a node, and the relations between them form intra-layer edges. The other kind of edge, the inter-layer connections, are set by the relationship between data from different information sources.

In this paper, we propose OD-HyperNet, a data-driven hyper-network model, to integrate the hyper-network model into the task of completing OD matrix. In real-world migration population estimating tasks, it works well in obtaining the inflows and outflows in most of the cities when such data of one or several cities are known. OD-HyperNet is composed of two single-layer networks, namely the population inflow network and the population outflow network. We use the inter-layer connections to link the nodes belonging to the same cities in the two-layer hyper-network, thereby connecting and aligning the two single-layer networks. Within OD-HyperNet, we utilize a cross-layer breadth-first traversal algorithm to estimate the population inflows and outflows. We also propose a practical completion procedure to lower the sparsity of the OD matrix.

Our experiment is based on the Baidu Migration dataset. It includes two parts: (1) the total number of people moving in and out of each region during a specific period, and (2) the inter-regional traffic flow, i.e., the number of people moving from one region to another. We evaluate the overall estimation accuracy base on an R-Squared metric, then show the influence of time and starting node. Results suggest that our model can complete the OD matrix very well and provide precise estimations with an accuracy of at least 0.975. To enhance the performance, we recommend to start traversal from cities with relatively high traffic and avoid tremendous fluctuations in data sequence.

2 Problem Definition

In this section, we stress the urgency of completing the origin-destination (OD) matrix, then present the assumptions and mathematical formulation for this problem.

2.1 Partially Observed Origin-Destination Matrix

A high-quality transit OD matrix acts as a solid foundation in many application problems, such as transportation planning [12], market evaluation [13], and epidemiological estimation [14]. Unfortunately, it is not always easy for people to obtain enough suitable data to build it.

There are two standard data providers for OD flow, but both of them have drawbacks. First, mobile network operators or cellphone APPs (e.g. China Mobile or Baidu Map) could generate real-time data by tracking their users. These data suffer from two aspects: they could only access their own users' moving trace, while the market share of them in different places is hard to analyze. Another provider, the government, could acquire more reliable data by surveys. However, such methods are expensive and sometimes result in old-fashioned data. Such a condition urges us to discover an appropriate method to complete the partially observed OD matrix.

In this paper, we focus on migration data from mobile carriers and cellphone APPs, for real-time information is more valuable to in driving decisions. To overcome the typical drawback of missing OD flows, we propose a data-driven modeling method, which can be applied to any dataset of this kind.

2.2 Basic Assumptions and Mathematical Formulation

Migration data provided by mobile carriers and cellphone APPs mainly fall into two categories: (1) the traffic flow between OD pairs, and (2) the total moving-in/out population of selected cities. Without loss of generality, we process these two types of data following three assumptions:

(1) *Assumption 1*: The market occupancy of a particular data provider in city i is α_i. Normally, when $i \neq j$, $\alpha_i \neq \alpha_j$.
(2) *Assumption 2*: The users of one data provider are uniformly distributed in the moving-in/out population. i.e., in city i, if its moving-in population is N_i^{in} and moving-out population is N_i^{out}, then the inflow/outflow users of the data provider are $\alpha_i N_i^{in}$ and $\alpha_i N_i^{out}$.
(3) *Assumption 3*: The inflow/outflow population occupies a relatively small portion of the total population in a city.

The analysis process runs as following: For the first type of data, we suppose that the number of people moving out of city i to city j is n_{ij}^{out}, the total number of people moving out of city i is N_i^{out}, then the proportion moving out of city i to city j is defined as $p_{ij}^{out} = n_{ij}^{out}/N_i^{out}$. Similarly, the proportion of newcomers of city j who come from city i is defined as $p_{ji}^{in} = n_{ji}^{in}/N_j^{in}$, where N_j^{in} represents the total number of migrants in city j and n_{ji}^{in} among them are from city i. Since $n_{ij}^{out} = n_{ji}^{in}$, $p_{ij}^{out} \cdot N_i^{out} = p_{ji}^{in} \cdot N_j^{in}$. Therefore, once the moving-in/out population in a city is known, the data of its related cities can be inferred. The first type of data is usually released as the outflow proportion matrix $\mathbf{P}^{out} = \{p_{ij}^{out}\}$ and the inflow proportion matrix $\mathbf{P}^{in} = \{p_{ji}^{in}\}$, which are estimated based on the user positioning data. In practice, these matrices are usually rather sparse, because there may not be any users travelling between for most of the city pairs.

The second type of data can be classified into two categories: (1) Total inflow and outflow population, i.e., N_i^{in} and N_i^{out}, which are released by the government. Such data is scarce but somewhat accurate; (2) Migration Index estimated by mobile network operators or Internet companies, which is a function of the moving-in/out users of a data provider. We indicate the moving-in and moving-out migration index of city i by $I_i^{in} = f(\alpha_i N_i^{in})$ and $I_i^{out} = f(\alpha_i N_i^{out})$ respectively.

The input of our model is the partially observed transport probability matrix $\mathbf{P}^{out}$ and $\mathbf{P}^{in}$. Without loss of generality, we assume that only the *Top-K* entities in each row of these two matrices are observed. That is, we only know the origins/destinations with the *Top-K* transport probabilities of each city. Besides, we also need at least one city whose accurate inflow or outflow population is known. That is, at least one N_i^{out} or N_i^{in} is needed.

The output of our model is the completed OD matrix $\{OD_{ij}\}_{|\mathbf{N}|\times|\mathbf{N}|}$, where $OD_{ij} = n_{ij}^{out}$ is the number of people who transport from city i to city j.

3 Model

3.1 OD-HyperNet: The Origin–Destination Hyper-Network Model

Based on the original definition of hyper-network [8, 10, 11], we propose the OD-HyperNet model to describe OD flow data.

A hyper-network $\mathbf{H}(\mathbf{B}(\mathbf{N}, \mathbf{E}), \mathbf{M}, \mathbf{R})$ is constructed from a base network $\mathbf{B}(\mathbf{N}, \mathbf{E})$, in which $\mathbf{N}$ represents the set of all members (nodes) and $\mathbf{E}$ represents the set of connections (edges). The set of single layer networks is denoted as $\mathbf{M}$, and the number of layers in the hyper-network is $|\mathbf{M}|$. In the behavior matrix $\mathbf{R}$, each element $\mathbf{R}(i, M)$ describes the connection pattern of node i in layer M.

Intuitively, the cities and the migration flows between them constitute a network, where cities are nodes, and the migration flows between the cities are edges. We build two migration networks, an inflow network, and an outflow network based on the previous definition.

In the outflow network, the weight of node i is set to N_i^{out}, which is the total number of people moving out of city i. The weight of directed edge (i, j) represents p_{ij}^{out}, which is the proportion of migrants from city i who move to city j. Similarly, in the inflow network, the weight of node j, N_j^{in}, is the total number of people moving into city j; the weight of directed edge (j, i),p_{ji}^{in}, is the proportion of the inflow population of city j coming from city i.

Notice that directed edge (i, j) in the outflow layer means that people move from city i to city j, while an edge (t, k) in the inflow layer shows the opposite direction, i.e., people move to city t from city k. Therefore, out-neighbors in the *outflow* layer represent the *origins* of inter-city trips, while the out-neighbors in the *inflow* layer are the *destinations*. Similarly, the *in-neighbors* in the *outflow* layer denote *destinations*, while *in-neighbors* in the *inflow* layer are *origins*. As available public data are often restricted to the *Top-K* origins/destinations of each city, the out-degree of each vertex in both layers is a constant, K. However, the in-degrees of each vertex in both layers can differ significantly. A high in-degree in the outflow layer denotes that the city acts as a frequent destination of the inter-city migration, while higher in-degree in the inflow layer means the city is more likely to be a common destination.

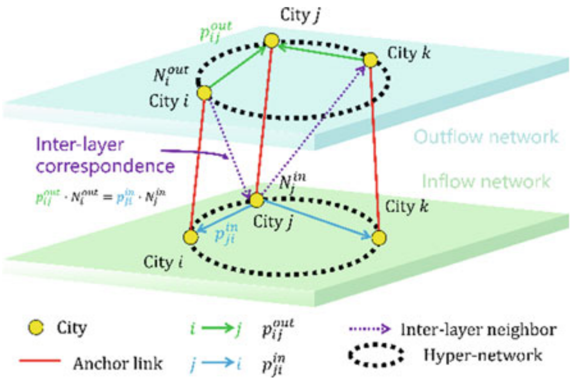


Fig. 1 An illustrative example for the inflow-outflow hyper-network

To construct the anchor links, we refer to the inter-layer correspondence between nodes (i.e., N_i^{out} and N_i^{in} belong to the same city i) and edge ($p_{ij}^{out} \cdot N_i^{out} = p_{ji}^{in} \cdot N_j^{in}$). Figure 1 gives an example of our proposed migration inflow-outflow hyper-network.

3.2 *Inferring Inflow/Outflow Population based on Cross-Layer Breadth-First Searches*

In most of the cases, we can calculate the proportion p_{ij}^{out} and p_{ji}^{in} from the data collected by mobile phone network operators. However, the precise numbers of inflow/outflow population (i.e.,N_i^{out} and N_j^{in}) are difficult to obtain. To conquer the problem, we first start from one city i with known inflow/outflow population, then iteratively derive the data of most cities based on $p_{ij}^{out} \cdot N_i^{out} = p_{ji}^{in} \cdot N_j^{in}$.

We apply the notion of interlayer neighbor to represent such relationship. City i and city j are interlayer neighbors if there is a pair of edges with opposite directions in two layers (e.g., directed edges $(i, j) \in \mathrm{E}_{out}$ and $(j, i) \in \mathrm{E}_{in}$, where E_{out} and E_{in} represent the set of edges of the outflow and inflow network, respectively). For one pair of inter-layer neighbors, if the total outflow population of city i is known, then the total inflow population of city j can be expressed as $N_j^{in} = p_{ij}^{out} \cdot N_i^{out} / p_{ji}^{in}$. Thus, the knowledge of only one or a few cities' inflows/outflows is enough to power the whole research. In Table 1, we propose a cross-layer breadth-first search (BFS) algorithm for a two-layer hyper-network as an example. Two critical catches here are (1) cross-layer BFS needs two queues to store the nodes of inflow and outflow layer, and (2) cross-layer BFS visits inter-layer neighbors iteratively rather than intra-layer neighbors.

Table 1 Cross-layer breadth-first search algorithm

```
Input:
Weights of all the edges in each layer
Weights of one or several nodes
Output:
Weights of all the nodes in each layer

1. Establish two queues: $Q_{in}$ and $Q_{out}$, enqueue nodes with known inflow data into $Q_{in}$, while nodes with known outflow data are loaded into $Q_{out}$. Here we suppose the weight of node $i$ in the outflow layer $N_i^{out}$ is known, thus we enqueue node $i$ into $Q_{out}$
2. While $Q_{out}$ is not empty, dequeue a node $i$ and visit its out-neighbors of the outflow layer, i.e., find the set of node $j$ where there exists $(i, j) \in E_{out}$
3. For each out-neighbor $j$, check whether edge $(j, i) \in E_{in}$ exists in the inflow layer. If such edge exists, we calculate the weight of node $j$ in the inflow layer by $N_j^{in} = p_{ij}^{out} \cdot N_i^{out} / p_{ji}^{in}$ and load node $j$ into $Q_{in}$
4. While $Q_{in}$ is not empty, dequeue a node $j$ and visit its out-neighbors of the inflow layer, i.e., find the set of node $k$ where there exists $(j, k) \in E_{in}$
5. For each out-neighbor $k$, check whether edge $(k, j) \in E_{out}$ exists in the outflow layer. If so, we calculate the weight of node in the outflow layer by $N_k^{out} = p_{jk}^{in} \cdot N_j^{in} / p_{kj}^{out}$ and load node $k$ into $Q_{out}$
6. Iteratively perform step (2) to step (5) until $Q_{in}$ and $Q_{out}$ are both empty
```

3.3 OD Matrices Completion

In the previous section, we obtain the weights (N_i^{out} and N_i^{in}) of each node i in both layers. Moreover, the weights of directed edges in each layer can be expressed as $p_{ij}^{out} = n_{ij}^{out}/N_i^{out}$ and $p_{ji}^{in} = n_{ji}^{in}/N_j^{in}$, where $n_{ij}^{out} = n_{ji}^{in}$ is the OD flow from city i to city j. It enable us to calculate the OD matrix $\{OD_{ij}\}_{|\mathbf{N}|\times|\mathbf{N}|}$, where $OD_{ij} = n_{ij}^{out} = n_{ji}^{in}$ and $|\mathbf{N}|$ is the number of the cities.

We can first calculate the OD matrix using either the inflow layer or the outflow layer. Since only the *Top-K* origins/destinations and their proportion are known, each row of the OD matrix only have K non-zero elements. Therefore, the number of the non-zero elements in the OD matrix is $\|\mathbf{OD}\|_0 = |\mathbf{N}|K$, where $\|\cdot\|_0$ denotes the L0-norm. Our goal is to complete the sparse matrix as much as possible, which can be regarded as a matrix completion problem [15]. Existing literatures reveal the spatial affinity feature of the OD matrix, that is, there are some rows similar to each other [16]. Therefore, we assume the OD matrices are low-rank.

However, compared with the general low-rank matrix completion problem, the OD matrix completion problem has a lot of intrinsic characteristics that can be leveraged. Much information can be provided by aggregating data from both layers in our OD-HyperNet model. Our OD matrix completion method has three steps.

(1) *Step 1*: We calculate two OD matrices based on the inflow layer and the outflow layer respectively. For the outflow layer, the OD matrix is calculated by $OD_{ij}^{out} = n_{ij}^{out} = p_{ij}^{out} \cdot N_i^{out}$. For the inflow layer, the OD matrix is calculated by $OD_{ij}^{in} = n_{ji}^{in} = p_{ji}^{in} \cdot N_j^{in}$.

(2) *Step 2*: We aggregate these two OD matrices based on the following equation,

$$OD_{ij} = \begin{cases} (OD_{ij}^{out} + OD_{ij}^{in})/2, \text{ if } p_{ij}^{out} \cdot p_{ji}^{in} \neq 0 \\ \max(OD_{ij}^{out}, OD_{ij}^{in}), \text{ if } p_{ij}^{out} \cdot p_{ji}^{in} = 0 \text{ and } p_{ij}^{out} \neq p_{ji}^{in} \\ 0, \text{ if } p_{ij}^{out} = p_{ji}^{in} = 0 \end{cases} \tag{1}$$

(3) *Step 3:* To further reduce the sparsity of the OD matrix obtained by Eq. (1), we apply the following low-rank matrix completion formulation,

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X}\|_* \\ &\text{subject to } X_{ij} = OD_{ij} \quad (i, j) \in \Omega \\ &\qquad \sum_{i=1}^{|\mathbf{N}|} X_{ij} = N_j^{in} \; j = 1, 2, \cdots |\mathbf{N}| \\ &\qquad \sum_{j=1}^{|\mathbf{N}|} X_{ij} = N_i^{out} \; i = 1, 2, \cdots |\mathbf{N}| \\ &\qquad X_{ij} \geq 0 \, i, j = 1, 2, \cdots |\mathbf{N}| \end{aligned} \tag{2}$$

where $\mathbf{X}$ is the decision variable (i.e., the completed matrix) and $\|\mathbf{X}\|_*$ is the nuclear norm of $\mathbf{X}$, which is applied to approximate the rank of $\mathbf{X}$. The set Ω contains the observed data. Program (2) aims at seeking the matrix $\mathbf{X}$ with the lowest rank that fits the observed data and the total inflow and outflow population.

Theoretically, Program (2) provides a proper formulation of the OD matrix completion problem. However, it may not be feasible in practice due to the bias in data. Thus, we provide an alternative formulation, Program (3), by relaxing some of the constraints.

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X}\|_* \\ &\text{subject to } X_{ij} = p_{ij}^{out} \quad (i, j) \in \Omega \\ &\quad \sum_{j=1}^{|\mathbf{N}|} X_{ij} = 1\, i = 1, 2, \cdots |\mathbf{N}| \\ &\qquad X_{ij} \geq 0\, i, j = 1, 2, \cdots |\mathbf{N}| \end{aligned} \tag{3}$$

In Program (3), $p_{ij}^{out} = OD_{ij}/N_i^{out}$, that is, we normalize each row of the OD matrix by N_i^{out}. Therefore, the decision variable X_{ij} is the proportion rather than the number of travelers. Therefore, the elements of the completed OD matrix is calculated X_{ij}. Program (3) is a constrained convex optimization problem, which can be solved by solvers such as CVXPY [17, 18].

4 Experiments

In this section, we apply the OD-HyperNet model to a case study using the Baidu Migration dataset, then validate the robustness and the accuracy of the OD flow estimation.

4.1 Data Description

The experiments are carried out based on the Baidu Migration dataset from Jan 1, 2020 to Jan 31, 2020. We collect the migration data of 352 cities. For each city, the dataset contains the following information: (1) *Top-100* origins or destinations with the proportion of daily population moving in/out of the given city, and (2) Baidu Migration Index that reflects the size of the population moving in or out the given city.

Table 2 shows the format of the origin-destination (OD) data. For example, the population who migrated from Wuhan to Xiaogan on Jan 23, 2020 accounts for 16.91% of the total outflow population of Wuhan. Likewise, the population that

Table 2 Examples of the origin-destination data

City A	City B	Date	Migration type	Proportion
Wuhan	Xiaogan	2020/01/23	Move-out	16.91%
Wuhan	Huanggang	2020/01/23	Move-out	14.12%
…	…	…	…	…
Shenzhen	Dongguan	2020/01/23	Move-in	16.77%
Shenzhen	Huizhou	2020/01/23	Move-in	11.84%
…	…	…	…	…

flows into Shenzhen from Dongguan on that day accounts for 16.77% of Shenzhen's total inflow population.

Table 3 lists the migration scale index data. It reads that the outflow population scale indexes of Wuhan and Shenzhen on Jan 23 were 11.14 and 17.78, respectively, which means that the total outflow population of Shenzhen on Jan 23 is 1.6 times higher than Wuhan.

In this paper, we first use the OD data to construct OD-HyperNet models for each day, then infer Baidu Migration Index based on the OD-HyperNet models. Assuming we only know one city's Baidu Migration Index, we validate the performance of our proposed inference method. The ground truth is the Baidu Migration Index of all the other cities.

4.2 *Robustness and Accuracy of Inflow/Outflow Population Estimation*

As is illustrated in Sect. 2, the accuracy of the OD matrix depends only on the accuracy of the estimated inflow and outflow population. Therefore, we evaluate the accuracy of the inflow/outflow population estimation. We also evaluate the robustness when starting from different vertices to perform cross-layer BFS. We demonstrate that only one starting vertex is needed, and our proposed method is not sensitive to the starting vertex. This finding is rather important because the set of candidates for the starting vertex (i.e., the cities whose inflow or outflow population is known) is rather limited in most of the practical applications.

Since the Baidu Migration dataset covers the *Top-100* origins/destinations of each city, the inter-layer and the intra-layer connections are rather dense in the OD-HyperNet models. However, this question remains to be determined: How many starting points does the cross-layer BFS need at least to infer the inflow and outflow population of all the other cities?

Since one city corresponds to two nodes (i.e., city i corresponds to N_i^{out} in the outflow layer and N_i^{in} in the inflow layer), there are 704 nodes in our proposed hyper-network model. We first choose each node as the starting point to perform the cross-layer BFS algorithm. It turns out that the cross-layer BFS starting from

Table 3 Examples of the migration index data

City A	Date	Migration type	Migration index
Wuhan	2020/01/23	Move-out	11.14
Wuhan	2020/01/23	move-in	1.75
…	…	…	…
Shenzhen	2020/01/23	Move-out	17.78
Shenzhen	2020/01/23	move-in	3.37
…	…	…	…

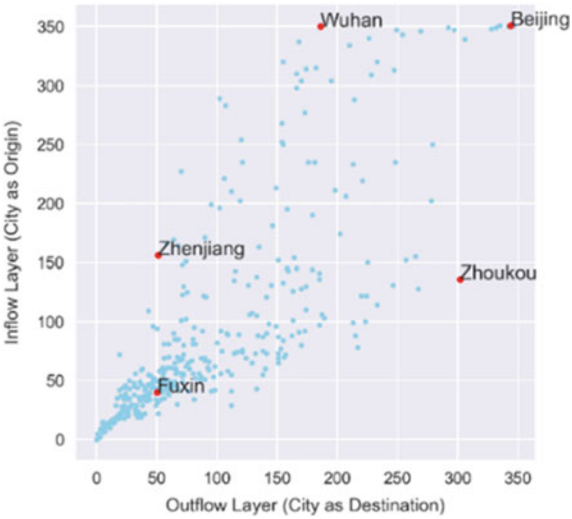


Fig. 2 The in-degrees in both layers and the starting points

any given node in the OD-HyperNet can reach all the other nodes. Therefore, we can select only one city, rather than a set of cities, as the starting point for each OD-HyperNet model to perform cross-layer BFS.

We choose 10 different nodes as the starting point of the cross-layer BFS algorithm and evaluate the overall estimation accuracy based on the R-Squared metric.

The starting points are chosen based on the in-degrees of the corresponding cities in both layers. Figure 2 shows a scatter plot, selecting the in-degrees of outflow layer and inflow layer to appear on the x-axis and y-axis, respectively. We choose 5 typical cities for our study and mark them in red in Fig. 3. For each city, we separately assume the inflow population N_i^{in} or outflow population N_i^{out} is known. We use the outflow migration index to represent N_i^{out}, and set N_i^{in} to the inflow migration index, thus obtaining 10 starting points to perform the cross-layer BFS.

We evaluate the overall estimation accuracy base on the R-Squared metric defined as Eq. (4). This metric is commonly used to evaluate the accuracy of the OD flow estimation [19]. The ground-truth of the inflow/outflow population of city i is denoted as N_i^{in} and N_i^{out}, the corresponding average values are expressed as $\overline{N_i^{in}}$ and $\overline{N_i^{in}}$. The estimation results are $\hat{N}_i^{in}$ and $\hat{N}_i^{out}$. Better estimations are indicated by higher R-Squared values [20].

$$\theta = 1 - \frac{\sum_{i=1}^{n}\left(N_i^{out} - \hat{N}_i^{out}\right)^2 + \sum_{i=1}^{n}\left(N_i^{in} - \hat{N}_i^{in}\right)^2}{\sum_{i=1}^{n}\left(N_i^{out} - \overline{N_i^{out}}\right)^2 + \sum_{i=1}^{n}\left(N_i^{in} - \overline{N_i^{in}}\right)^2} \tag{4}$$

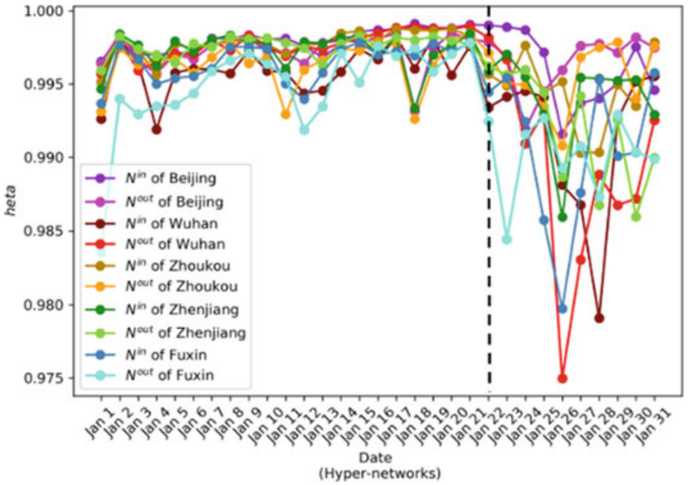


Fig. 3 Inference accuracy at varying starting points

Figure 3 illustrates the R-Squared values for the inference results of the cross-layer BFS starting from N_s^{out} or N_s^{in}, where s is one of the 5 cities.

There are three key observations of Fig. 3. First, the overall inference results are rather accurate, and the performance is relatively robust to different choices of the starting point. Second, choosing vertex with higher in-degrees improves the accuracy of the inference regardless of the layer. This indicates that transportation centers are more effective as the starting points. Third, the inference accuracy is rather stable from Jan 1 to Jan 22. However, the overall inference accuracy dramatically fluctuates after Jan 23. The reason might be many cities have had issued traffic restricting policies after Jan 23. Since Baidu may calculate the original data (e.g., N_i^{out} and p_{ij}^{out}) based on specific traveling pattern, significant changes in traveling pattern can lead to larger error and instability.

We further calculate the estimation error of each value based on $E_i^{io} = (N_i^{io} - \hat{N}_i^{io})/N_i^{io}$, where N_i^{io} represents either N_i^{in} or N_i^{out} of the city i. Figure 4 shows that most of $E_i^{io} \in (-0.2, 0.2)$, that is, most of the inference error is within the range of $(-0.2N_i^{io}, 0.2N_i^{io})$. We compare the distribution of E_i^{io} for the cross-layer BFS algorithm starting from 4 different vertices (i.e., the inflow/outflow population of Beijing and Fuxin) in Fig. 5. This agrees with our previous results that starting from N_i^{in} or N_i^{out} of big cities can be more accurate.

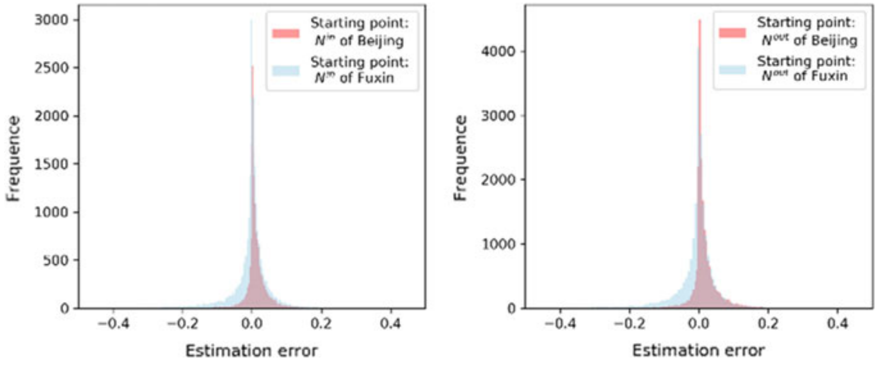


Fig. 4 Distribution of error at varying starting points

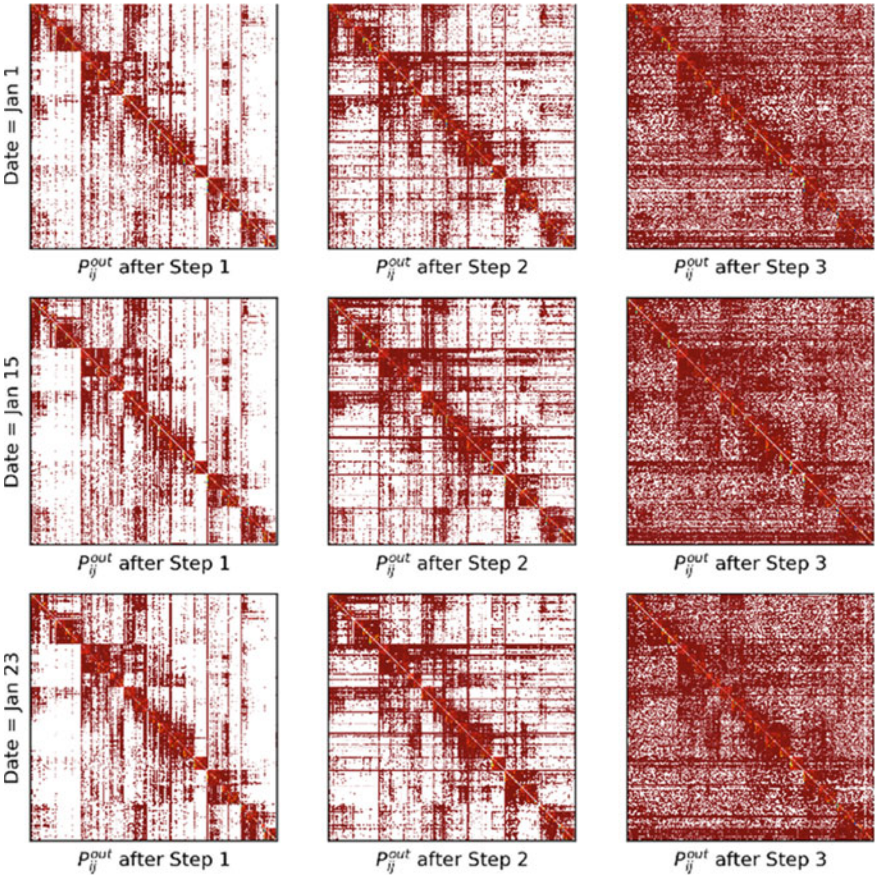


Fig. 5 The matrix completion results of three steps

4.3 Completing OD Matrix

In this part, we evaluate the model's performance in completing the OD matrices. Since we do not have the ground truth of the missing data, we only compare the sparsity of the matrix after Step (1)–(3).

Figure 5 gives examples of the OD matrices after each step. Each row represents the OD matrices at one day. The three columns are the OD matrices after Step (1)–(3), respectively. For better visualization, we normalize each column of the OD matrices by the total outflow population of the corresponding city, that is, we obtain $p_{ij}^{out} = OD_{ij}/N_i^{out}$ and visualize the matrix $\left\{p_{ij}^{out}\right\}_{|\mathbf{N}|\times|\mathbf{N}|}$.

Figure 5 demonstrates that our OD matrix completion procedure gradually decreases the sparsity of the OD matrices step by step to ultimately form a rather dense OD matrix. Interestingly, it can also be noticed that there are some time-invariant patterns of the matrices in each column.

5 Conclusion

Accurate tracking of population flow is crucial in the prediction and containment of possible infections during the current COVID-19 epidemic. However, real-world data are not accessible in many cases, which may result in sparse OD matrices. This paper solves the problem by incorporating hyper-network model with data-driven method. Our model is able to interpolate and infer missing pieces of data and provides high estimation accuracy and matrix completeness.

Acknowledgements This research was funded by the National Natural Science Foundation of China (Grant No. 71971127) and the Hylink Digital Solutions Co., Ltd. (120500002).

References

1. D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena. Science **342**(6164), 1337–1342 (2013)
2. D. Taylor et al., Topological data analysis of contagion maps for examining spreading processes on networks. Nat. Commun. **6**(1), 1–11 (2015)
3. J.S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, N.A. Christakis, Population flow drives spatio-temporal distribution of COVID-19 in China. Nature, 1–5 (2020)
4. D.E. Low, New approach to transportation systems modeling. Traffic Q. **26**(3) (1972)
5. O. Tamin, L. Willumsen, Transport demand model estimation from traffic counts. Transportation **16**(1), 3–26 (1989)
6. M.-H. Wang, S.D. Schrock, N. Vander Broek, T. Mulinazzi, Estimating dynamic origin-destination data and travel demand using cell phone network data. Int. J. Intell. Transp. Syst. Res. **11**(2), 76–86 (2013)

7. L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, L. Damas, Time-evolving OD matrix estimation using high-speed GPS data streams. Expert Syst. Appl. **44**, 275–288 (2016)
8. W.K.V. Chan, C. Hsu, Service scaling on hyper-networks. Serv. Sci. **1**(1), 17–31 (2009)
9. W.K.V. Chan, C. Hsu, How hyper-network analysis helps understand human networks? Serv. Sci. **2**(4), 270–280 (2010)
10. W.K.V. Chan, C. Hsu, Service value networks: humans hypernetwork to cocreate value. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **42**(4), 802–813 (2012)
11. W.K.V. Chan C. Hsu, When human networks collide: the degree distributions of hyper-networks. IIE Tran. **47**(9), 929–942 (2015)
12. R. Borndörfer, M. Grötschel, M.E. Pfetsch, A column-generation approach to line planning in public transport. Transp. Sci. **41**(1), 123–132 (2007)
13. T. Li, A demand estimator based on a nested logit model. Transp. Sci. **51**(3), 918–930 (2017)
14. Z. Cao et al., Incorporating human movement data to improve epidemiological estimates for 2019-nCoV, medRxiv (2020)
15. E.J. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717 (2009)
16. H. Zhou, D. Zhang, K. Xie, Accurate traffic matrix completion based on multi-Gaussian models. Comput. Commun. **102**, 165–176 (2017)
17. S. Diamond, S. Boyd, CVXPY: a Python-embedded modeling language for convex optimization. J. Mach. Learn. Res. **17**(1), 2909–2913 (2016)
18. A. Agrawal, R. Verschueren, S. Diamond, S. Boyd, A rewriting system for convex optimization problems. J. Control Decis. **5**(1), 42–60 (2018)
19. A. Tavassoli, A. Alsger, M. Hickman, M. Mesbah, How close the models are to the reality? Comparison of transit origin-destination estimates with automatic fare collection data, in *Proceedings of the 38th Australasian Transport Research Forum (ATRF)*, pp. 1–15 (2016)
20. X. Liu, P. Van Hentenryck, X. Zhao, Optimization Models for Estimating Transit Network Origin-Destination Flows with AVL/APC Data, arXiv preprint arXiv:1911.05777 (2019)

The Supervision Strategy Based on Evolutionary Game Between Government and Sharing Economy Enterprises of Beijing

Dandan Li and Xiaochuan Li

Abstract In recent years, sharing economy is developing rapidly. The sharing economy enterprises in Beijing have the characteristics of being cross-industry, networked, large in number, and wide-ranging. So they put forward new requirements on the current regulatory model. In this paper, the dynamic evolutionary game model between government and sharing economy enterprises of Beijing is established, and the stability of the game is discussed, in order to provide some supervision strategy for government of Beijing. How to choose the strategy for both government and sharing economy enterprises of Beijing? The model proposed in this paper could not only achieve the encourage innovation, but also could standardized the development of sharing economy enterprises in Beijing. The game evolution results show that the third-party reporting rate, penalties for violations, and the success rate of government supervision have a great influence on the Beijing government's strategic choice. This paper provides the theoretical basis for the strategy selection and policy formulation in the period of sharing economy. At the same time, it could provide some theoretical basis for government of Beijing at governance to sharing economy enterprises, and the developing strategy of enterprises based on government of Beijing.

Keywords Dynamic evolutionary game theory · Sharing economy enterprise · Government supervision

1 Introduction

With sharing economy development and social progress, science and technology play a more and more important role in it, especially in super large cities like Beijing. There are many kinds of sharing economy enterprises such us food, transport, bikes, houses,

D. Li (✉)
School of Management, Beijing Union University, Beijing 100101, China
e-mail: gldandanli@buu.edu.cn

X. Li
Information Office, Capital University of Economics and Business, Beijing 100070, China
e-mail: xc_li@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_25

even knowledge and so on. So the government supervision is more difficult and complex for Beijing than other regions. The catalytic effect of technological progress on economic growth is more and more obvious, especially in the advanced nowadays of knowledge economy. Technological progress in all dimensions is going beyond the past, and it has played a huge role in promoting economy. Due to the correlation between economic growth and technological growth, economic growth among the countries is not balanced in all over the world. The Beijing government and sharing economy enterprises play the game constantly in their development process. On the one hand, the guidance of policies is conducive to the technological sharing economy enterprises. On the other hand, the government's role of Beijing in impeding the cost of technological developing of enterprises increases. The development of sharing economy enterprises has its particularity, which belongs to the initial stage. Many legislation and policies are not perfect, so it is necessary to explore and standardize at the same time in Beijing.

Sharing economy enterprise belongs to technology dependence. Government of Beijing plays the role of supervision on its developing. According to Product Cycle Theory of Vernon [1], the relationship between government and enterprises is as follows: government, because of it owns research and development (R&D) resources police, whether block or not is very important to sharing economy enterprises. Some new products become easier to be imitated because they are standardized slowly. The technology spill-overs and technology diffusion are obvious. The developing enterprises started to follow the leading ones through technology imitation, which mainly imitates and produces the standard products of the leading region through labor cost advantage. So, government supervision could adjust including supervision technology, supervision method, supervision platform and supervision means. Meanwhile, how to strike a balance between encouraging innovation and standardized development, analyze the impact of third-party reporting rates, violation penalties, and government supervision success rates on the government's strategic choices, and provide a basis and reference for the government's decision-making. The "northern innovation-southern imitation" model (North–South Model) followed by later scholars of the technological developing, such as Krugman [2], Dollar [3], Jensen et al. [4], Grossman et al. [5], Barro et al. Wait. In this paper, it is used to study Beijing government and sharing economy enterprises about their evolutionary game.

In order to study the relationship of supervision between Beijing government and sharing economy enterprises, literature research is increased not only including Beijing, but also the other general business and governments. The research on the development of sharing economy companies can learn from domestic and foreign scholars' related research on technological catching-up. Alexander et al. [6] suggested that relatively developing region could absorb the existing technologies (at a relatively low cost). It should have a faster growth rate, and rely on technology spill-overs to narrow the gap with leading region and achieve technology advancement. Such as the "leap-frogging" encountered by the New Growth Theory [10]. However, this phenomenon of "late-onset technological superiority" did not occur in the vast majority of developing region, but also emerged as a "paradox of late-onset superiority,” that is, The technological gap between developed leading regions and some developing regions is getting bigger and bigger. In fact, when Alexander et al. [6] proposed the theory of “ Latecomer technological advantages”, he pointed out that it takes huge costs for developing region to imitate technology simultaneously, such as building a system compatible, facilities and policies with new technologies. Therefore, many scholars use it to explain the “late-onset superiority paradox” as well as “the suitability of technology introduction”, the adaptability of factor endowments, the efficiency of financial markets, the industry-related effects and the obstacles Technical absorption of various types of technical barriers and other factors [7]. Other scholars think that the technological capability (including learning ability and absorptive capacity) of developing region is the key to narrow the technological gap and to realize technological developing [8–11]. It mainly focuses on whether it could use the “latecomer technological advantages. In other words, when a region has strong technical capability, its ability to digest and absorb technology spill-overs is strong too. Based on the research and conclusions of the new growth theory, under the condition that technological progress is still an indisputably engine of economic growth [12]. Barro et al. [13] think that the reason for the lack of economic convergence in the technologically backward countries is the cost of technology imitation. In the early stage of technology imitation, the backward countries (regions) in technology imitation had low cost of imitation, so the economy grew rapidly and the technology level continued to rise. When the gap between the technological level and the leading countries (regions) in the technology shrank to a certain extent, the difficulty of imitation and the cost will be higher and higher, and some of the core technologies in the leading countries and regions of technology are hard to be imitated. But Zou [14] thinks that the concept from Barro etc. of imitation costs is not clear. Sheng [15] thinks factor endowments determine the technical ability. And both of them determine the ability of technological developing. Huang et al. [16] discuss the technology gap between China and the United States from the perspectives of short-term and long-term respectively. These studies discuss the factors that influence the success of technology developing from different perspectives. However, there is no clear definition of the relationship between the developing region and leading region, and technology developing is taken as a single acting of developing region, the interaction between two sides is ignored. Previous scholars seldom discussed how the government should promote the catching up and catching up development of the sharing economy enterprises based on their weak ability and lack of power. In our opinion, the same situation is also applicable between technology leading enterprises and technology backward enterprises.

Beijing is one of the highest technology centers in China. So as the sharing economy enterprises in Beijing, their technological catching-up has largely affected the development of sharing economy in China. The Beijing government plays a role of policy adjustment among the enterprises of sharing economy. The key of catching-up is of technology, and the core of technology catching-up is the policy application. Whether Beijing government supports their catching-up to other leading enterprises in developed country, for instance at tax. It is very import to their strategy. Therefore, in order to improve the R&D efficiency of sharing economy enterprises as the starting point, and to improve the driving force and ability of technology catching-up, constructs an evolutionary on the theory of policy tools, this paper constructs a policy framework for the Beijing government to support the enterprises of sharing economy to catch up, and discusses the policy tools for the Beijing government to support the enterprises and realize the catching-up.

2 Game Matrix

2.1 Game Matrix

Due to the asymmetry of information, both Beijing government and sharing economy enterprises have no idea what tactics they will adopt when they make decisions. So, as a game of bounded rationality, there are two choices of enterprises: one strategy is to catch-up (developing strategy), the other is not to catch-up (non-developing strategy). Similarly, there are two choices of Beijing government, too: one is to block the developing strategy, that is the not support strategy (short for block strategy), the other is not to (short for non-block strategy), that is Beijing government supports the enterprises to catch up.

(1) If the block strategy is adopted by sharing economy enterprises, at the same time, the non-developing strategy is adopted by Beijing government, both of their innovation income is 0.

(2) If the block strategy is adopted by Beijing government, and developing strategy is adopted by Beijing sharing economy enterprises (maybe because of policy support from domestic government, economic development orientation, endogenous efforts), then, it is possible for sharing economy enterprises to catch-up (assume that the probability is d_s^1, profit is Π_s^1), or not (assume that the probability is $1 - d_s$, profit is L_s). The cost increase rate which cost because of the blocking strategy is e_s^{l}, and the profit of sharing economy enterprises is $d_s^1 \Pi_s^1 \left(1 - e_s^1\right) - (1 - d_s^1)\, L_s$, government is $\left(1 - d_s^1\right) \Pi_s^1 e_s^1$.

(3) The non-block strategy is adopted by Beijing government, and developing strategy is adopted by sharing economy enterprises. At this moment, the developing strategy of sharing economy enterprises is supported by increasing spill-over and key technology of Beijing government. Then, it is possible for sharing economy enterprises to catch-up (assume that the probability is d_s^2, profit is Π_s^2), or not (assume that the probability is $1 - d_s^2$, profit is L_s). The cost increase rate which cost because of the blocking strategy is e_s^2, and the profit of sharing economy enterprises is $d_s^2 \Pi_s^2 (1 - e_s^2) - (1 - \mathrm{d}_s{}^2)\, L_s + S$, Beijing government is $(1 - d_s{}^2) \Pi_s^2 e_s^2 - S$. Because when spill-over is creasing, market share and profit of this market have to be lost, so S is used to represent it.

(4) If the non-block strategy is adopted by Beijing government, and non-developing strategy is adopted by sharing economy enterprises, then, the profit of sharing economy enterprises is S, government is $-S$.

Table 1 Game gain matrix

Beijing Government	Sharing economy enterprises		
		Developing strategy	Non-developing strategy
	Non-block strategy	$\left(1-d_s^2\right)\Pi_s^2 e_s^2 - S, d_s^2\Pi_s^2\left(1-e_s^2\right) - \left(1-d_s^2\right)L_s + S$	$-S, S$
	Block strategy	$\left(1-d_s^1\right)\Pi_s^1 e_s^1, d_s^1\Pi_s^1\left(1-e_s^1\right) - \left(1-d_s^1\right)L_s$	0, 0

Game gain matrix of Beijing government and sharing economy enterprise is built based on the above analysis. It is shown in Table 1.

Firstly, if other conditions are the same (or other conditions are not considered), then d_s^2 is bigger than d_s^1, that is $d_s^2 > d_s^1$.

Secondly, when block strategy is adopted by sharing economy enterprise, the cost increasing will be bigger than non-block strategy is adopted, that is $e_s^1 > e_s^2$.

Thirdly, when non-block strategy is adopted by Beijing government, the profit of sharing economy enterprise is bigger than block one, that is $\Pi_s^2 > \Pi_s^1$.

2.2 Evolutionary Game Model of Technological Developing Behavior

We suppose that the probability of developing strategy is adopted by sharing economy enterprises is x, and non-catching strategy is adopted is $(1-x)$. Non-block strategy is adopted by Beijing government is y, and block strategy is adopted is $(1-y)$. Copy dynamic equation is used to analyze the strategy adjustment between developing and Beijing government during the developing.

2.3 Replicated Dynamic Equation Sharing Economy Enterprises

According to the above analysis, the expected return of the sharing economy enterprises when adopting the catching-up strategy is shown in Formula (1).

$$u_{se} = y\left[d_s^2\Pi_s^2\left(1-e_s^2\right) - \left(1-d_s^2\right)L_s + S\right] + (1-y)\left[d_s^1\Pi_s^1\left(1-e_s^1\right) - \left(1-d_s^1\right)L_s\right] \quad (1)$$

Otherwise, the expected return is

$$u_{sn} = y \cdot S + (1 - y) \cdot 0 = yS \tag{2}$$

So, the average return is

$$\overline{u}_s = xu_{se} + (1 - x)u_{sn} \tag{3}$$

The replication dynamic equation of sharing economy enterprises F is

$$F_s(x) = x(u_{se} - u_s) = x(1 - x)(u_{se} - u_{sn}) \tag{4}$$

2.4 Replicated Dynamic Equation of Government

The expected return of the Beijing government when adopting the non-block strategy is

$$u_{ne} = x[(1 - d_s^2)\Pi_s^2 e_s^2 - L_s - S] + (1 - x)(-S) = x(1 - d_s^2)\,\Pi_s^2 e_s^2 - S \tag{5}$$

The expected return of Beijing government when adopting block strategy is

$$u_{nn} = x[(1 - d_s^1)\,\Pi_s^1 e_s^1] \tag{6}$$

The average return is

$$\overline{u}_n = yu_{ne} + (1 - y)u_{nn} \tag{7}$$

So, the replication dynamic equation of sharing economy enterprises is

$$F_n(y) = y(u_{ne} - \overline{u}_n) = y(1 - y)(u_{ne} - u_{nn}) \tag{8}$$

2.5 Dynamic Evolution of Game Model

According to the stability theorem of the differential equation, the derivative of the function F must be less than zero at the steady state. It is represented by the phase map of the replicated dynamic equation, which is the point where the horizontal axis intersects and the slope is negative. When $F_s(x) = 0$, $\mathrm{F}_n(y) = 0$, then:

$$\begin{cases} F_s(x) = x(1-x)(u_{se} - u_{sn}) = 0 \\ F_n(y) = y(1-y)(u_{ne} - u_{nn}) = 0 \end{cases} \tag{9}$$

Equilibrium point (x^*, y^*) of Beijing government and sharing economy enterprises is (0,0), (1,0), (0,1), (1,1) and (a,b), where

$$a = \frac{S}{(1-d_s^2)\pi_s^2 e_s^2 - (1-d_s^1)\pi_s^1 e_s^1}$$

$$b = -\frac{d_s^1 \pi_s^1 (1-e_s^1) - (1-d_s^1)L_s}{d_s^2 \pi_s^2 (1-e_s^1) - (1-d_s^2)L_s - [d_s^1 \pi_s^1 (1-e_s^1) - (1-d_s^1)L_s]}$$

And $d_s^1 \pi_s^1 (1-e_s^1) - (1-d_s^1)L_s > 0$.

Then, steady state neighborhood stability is discussed. Are these points ESS (evolutionary stable strategy) if there is minor deviation disturbance?

The steady state of x and y, could be discussed according to different a and b.

There are three situations such as 0 < a < 1 and b > 1, a > 1 and b > 1, a > 1and 0 < b < 1, the ESS should be (0,0). In other hand, if a < 0 and b > 1, the ESS should be (1,0). At last, when {0 < a < 1 and b < 0} or {a < 0, 0 < b < 1} or {a < 0 or a > 1 and b < 0}, the ESS should be (1,1).

3 Analysis of Game Evolution

There are three motion tracks which are track 1, track 2 and track 3. Through analysis we can see that, for sharing economy enterprises, the game evolution path should from track 1, track 2 to track 3. That is from (block strategy, non-catching up) to (non-block strategy, catching up). It is proved stable, and it is in line with the objective laws of today's social development. The conclusion is as following:

(1) When $S > (1-d_s^2)\pi_s^2 e_s^2 - (1-d_s^1)\pi_s^1 e_s^1$, that is $(1-d_s^2)\pi_s^2 e_s^2 - S < (1-d_s^1)\pi_s^1 e_s^1)$ and $d_s^2 \pi_s^2 (1-e_s^2) - (1-d_s^2)L_s < d_s^1 \pi_s^1 (1-e_s^1) - (1-d_s^1)L_s$, the profit of non-block strategy for sharing economy enterprises is lower than block one. So, block strategy should be adopted by Beijing government. In the process of catching-up for sharing economy enterprises, because of technology gap, low technology spill-over, high cost of technological innovation, and stop of Beijing government, the expected return is less than zero, although the developing strategy might be adopted by sharing economy enterprises. No one wants to catch up because the profit is negative number. So, catching up thinking and power will be lost and lead to non-catching up at last. Then, the ESS is (0,0). Similar, when the profit of non-block strategy for Beijing government and enterprises of Beijing government is lower than block one, the non-block strategy will

be adopted. The expected return is less than zero. It may be due to poor management, unfamiliar with technology spillover applications, fierce internal competition, imperfect supporting facilities, lack of technical staff for enterprises and so on. Then developing strategy will not be adopted by sharing economy enterprises. There is no return on technology spill-over and diffusion of government, the expected return will become negative number. The power of non-block will be lost, so the evolution of both game sides ESS is (0,0). If technology spill-over is getting smaller, then technology gap will be getting bigger. The speed of standardization for creative products and technology will be getting slow because of the above situation. Even when the technology is known by public and it is easy to be copied, but the sharing economy enterprises still cannot master it, maybe because the technology gap is too big. Furthermore, Beijing government could not to do support for new technology research and sharing economy enterprises is still learning old technology. This situation is not good for the development of the global economy and the emergence of new technologies. In other words, it will slow down the economic development of the whole market.

So, when the above situation appearance, support strategy should be adopted by Beijing government to help sharing economy enterprises to catch-up. Technology spill-over should be increased and the probability of successful to R&D which is d_s^2 is heightened.

(2) When the average return of Beijing government who adopts block strategy is larger than non-block ones. At the same time, the profit of sharing economy enterprises who adopts developing strategy is larger than non-catching up ones, and both of them are bigger than zero. So there is power for them to adopt developing strategy, and there is no power for them to adopt non-block strategy. But even so, there are some other reasons for sharing economy enterprises to adopt developing strategy, such as the support of Beijing government (police, money, people, place, etc.), high level of enterprises technology stock, lower cost of their own and so on. They are going to occupy market share and even technology dominance in emerging technology markets. In this case, the cost of block strategy is increasing and it is insufficient investment in new technologies. Innovative efficiency is reduced. At last, new technology is successful created by sharing economy enterprises (such as leaping frog). So, for Beijing government, the block strategy should not be adopted. Conversely, technology spill-over should be expanded appropriately, and so as technology guidance. The copy and learning advanced technology for Beijing government should be supported and that is good for technology standardization. The developing will be shown soon.

(3) When the expected return for sharing economy enterprises, who adopted developing strategy, is bigger than zero, and profit of non-block strategy is bigger than block one. There is power for sharing economy enterprises to catch up, and so as Beijing government to support (that is non-block). At last, (non-block, developing) will be the best strategy for both of them. It is the ideal state to keep stable for marketing. Most of the profit of sharing economy enterprises is from standard products. As the grown of profit, more and more capital could

be used for R&D. Meanwhile, technology spill-over and key technology guidance is coming from the help of Beijing government. All of them are good for catching-up. R&D staff from enterprises has to do more and more research for new technology, which is speeder the development of economy and innovation. Moreover, economy technology of the whole society is faster too. The virtuous cycle is formed.

4 Conclusion

In our opinion, non-block strategy of Beijing government is relative in practice, and the core technology is not open fully.

From the above game analysis of the Beijing government and sharing economy enterprises, it can be seen that the equilibrium point of the game evolution between them is adjusted and dynamically changed according to the development of policies and catching-up strategies. With the development of the sharing economy, Beijing government has become more and more flexible in the form of its supervision. This article proposes to improve the government supervision mechanism that is compatible with the sub-economy from a different perspective. In terms of regulatory technology, make full use of technologies that combine cloud computing, Internet of Things, and big data; in terms of regulatory methods, it is recommended that the Beijing government standardize market access mechanisms to keep corporate information confidential, and at the same time disclose related corporate deficiencies in a timely manner Information, improve the public's right to know and supervise; in terms of supervision methods, sound process supervision includes full supervision before, during and after the incident, and integrated online and offline supervision; finally, build a network guided by the Beijing government -The transaction supervision service platform releases and synchronizes the above technologies, methods and methods on an online platform to realize the innovative supervision and governance of the sharing economy enterprises by the Beijing government. So, both of Beijing government and sharing economy enterprises should take positive measure to form this stable status, in order to realize virtuous cycle of the whole society.

Acknowledgements This work was supported by Beijing Municipal Education Commission project (SM201811417008).

References

1. P.R. Krugman, A model of innovation, technology transfer, and the world distribution of income. J. Polit. Econ. **87**(1), 253–266 (1979)
2. D. Dollar, Technological innovations, capital mobility, and the product cycle in north-south trade. Am. Econ. Rev. **76**(1), 177–190 (1986)

3. R. Jensen, M. Thursby, A strategic approach to the product life cycle. J. Int. Econ. **21**(1), 269–284
4. G.M. Grossman, E. Helpman, Quality ladders and product cycles. Quart. J. Econ. **106**(1), 557–586 (1991)
5. R.J. Barro, X. Sala-i-martin, Technological diffusion, convergence, and growth. J. Econ. Growth **2**(1), 1–27
6. G. Alexander, *Economic Backwardness in Historical Perspective* (Harvard University, Boston, 1962)
7. E. Brezis, P. Krugman, D. Tsiddion, Leapfrogging: a theory of cycles in national technological leadership. Am. Econ. Rev. **83**(1), 1211–1219 (1993)
8. D.D. Li, L.Y. Ma, Technological Capacity, path dependence and technology developing. Sci. Technol. Progr. Policy **33**(17), 20–24 (2016)
9. Y. OuYang, Y.C. Sheng, Technology gap, technological capacity and technology developing in developing regions. China Soft Sci. **2**, 153–160 (2008)
10. Y.C. Sheng, Factor technical ability and the evolution of the technical developing. Fin. Trade Res. **2**, 46–54 (2010)
11. A. Botta, A structuralist north–south model on structural change, economic growth and developing. Struct. Change Econ. Dyn. **20**(1), 61–73 (2009)
12. B. Verspagen, A new empirical approach to catching up or falling behind. Struct. Change Econ. Dyn. **2**(2), 359–380 (1991)
13. R.J. Barro, X. Sala-i-martin, *Economic Growth* (McGraw-Hill, Inc., New York, 1995)
14. W. Zou, Q. Dai, Technological Imitation, human capital accumulation and economic developing. Soc. Sci. China **5**, 26–37 (2003)
15. Y.C. Sheng, *The Factor Endowments, the Technical Capability and the Technical Catching Up of the Developing Economy* (Hunan University, Changsha, 2008)
16. Y.C. Huang, M.L. Chen, X.L. Chen, Has the technology gap between china and the usa been decreased?—empirical analysis of China and United States in 1996–2012. Sci. Technol. Progr. Policy **15**, 1–8 (2016)

Research on Learning Path Recommendation in Intelligent Learning

Wenjing Dong and Xuedong Chen

Abstract This paper aims at the stability problem of Ant Colony Algorithm in learning path recommendation. First of all,explore how to select intelligent algorithm to realize learning path recommendation in intelligent learning; then, based on Ant Colony Algorithm, the basic information, learning style and knowledge level of learners and the expression form and difficulty coefficient of learning objects are considered to recommend learning path; finally, the value of volatilization factor ρ of Ant Colony Algorithm is adjusted, the simulation experiment is carried out by using control variable method, and the best volatilization factor is selected to improve the stability of the algorithm.

Keywords Intelligent learning · Ant colony algorithm · Learning path recommendation

1 Introduction

Driven by the tide of high integration of information technology and education development, a new type of intelligent learning model has been produced. One of the core services of intelligent learning system is to recommend the best learning path for learners according to their individualized characteristics. According to the learning characteristics and learning path of the target users, intelligent algorithm is used to select a suitable learning path for the target users, so as to help the target users achieve the established learning goals. In this paper, the learner's learning characteristics are taken as the input of the algorithm, the learning path is taken as the output of the algorithm, and the recommendation strategy based on the learning path is realized in combination with the learner's review characteristics and existing problems and

W. Dong (✉) · X. Chen
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: wjdongwj@163.com

X. Chen
e-mail: xdchen@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_26

so on. Only in this way can realize the selection and optimization of the intelligent algorithm, and realizes the intelligent learning.

2 Literature Review of Learning Path Recommendation

2.1 Influencing Factors of Learning Path Recommendation

In the aspect of the influence factors of the learning path recommendation, in 1996, 《Methods and techniques of adaptive hypermedia》published by Brusilovsky [1] described adaptive hypermedia as a new research direction in the field of adaptive and interfaces based user model, This is an enlightening paper in the field of learning path recommendation in intelligent learning system. An adaptive hypermedia system models an individual user and applies it to that user, for example, adjusting the content of a hypermedia interface to the user's knowledge and goals, or suggesting the most relevant links, which is constantly improved and supplemented by later scholars. However, the model is limited to the knowledge level and learning goal of the target user, and does not consider the learning style of the individual user. Chen [2] based on genetic algorithm, at the same time, in learning path recommendation, considering the matching degree between the difficulty of learning objects and the level of learners' knowledge, and considering the continuity of the concept of learning path. Personalized ranking is carried out in the network learning system, it's a way better than free browsing. With the deepening of the research on educational theory and practice, relevant scholars have pointed out that the learning process should not only consider the learners 'learning goals and knowledge level, but also the learners' learning style, for example, as Judy [3] proposes an innovative adaptive learning method according to the two main sources of personalized information learning behavior and personal learning style, which is helpful to improve students' learning achievement and learning efficiency. In an adaptive learning system, it is a very important step to find the learning style of the learners. Regarding this problem, Chang [4] proposed a learning style classification mechanism, improved k-nearest neighbor classification algorithm and combined with genetic algorithm, classify and identify the learning styles of the students. In recent years, with the in-depth research on learning path recommendation, Zhao [5] added the dimension of learning state to learning path recommendation and divided it into initial state, adaptive state, change state and end state, which is used to depict learners' learning situation in current learning path.

Through the summary of the above-mentioned research process, it is found that the influencing factors of learning path recommendation mainly include four aspects: learners' learning objectives, knowledge level, learning style and learning state, but the dimension of learning state conflicts with the personalized recommendation service in intelligent learning system, and the original intention of learning path recommendation is to adapt to the personalized characteristics of learners. This paper only considers three factors: learning goal, knowledge level and learning style.

2.2 Intelligent Algorithm of Learning Path Recommendation

(1) *The research process of intelligent algorithm:* Bert [6] based on self-organization theory and the method of group intelligence to realize the recommendation of learning path, through recording, processing and presenting the behavior of group learners, to establish a feedback loop, inform learners of successful approaches to achieving their learning goals. This method does not take into account the learners' learning style and knowledge level. It is based on the learners' learning goal directly, and then uses the information of the learning group to realize the recommendation. The accuracy of this method is poor. In the literature study of the influence factors of the recommended learning path, it is mentioned that the learners' knowledge level and learning style play an important role in improving the learners' learning efficiency. As in Yang [7], an attribute ant algorithm is proposed, which takes learners' knowledge level, learning style and the relationship among the attributes into account. An adaptive learning rule is proposed to define learning paths with high probability. The Algorithm only considers the frequency of learning path used by the learning group, and ignores the feedback of learning effect. On the basis of the above, in Cheng [8], the improvement of Group Intelligence Algorithm is mainly shown in two aspects: In the source of pheromone, the scope is limited to the learners with similar learning styles; in the content of Pheromone, it includes not only learners' evaluation of the learning object, but also learners' evaluation of the whole learning path. Many online learning systems require learners to evaluate the learning object after learning, this will increase the precision of recommendation decisions. With the concept of micro-learning proposed, combining it with learning path recommendation, the recommendation granularity changes from an entire path to a learning unit. Therefore, Zhao [9] suggests an ant colony algorithm to adjust the pheromone concentration step by step. The algorithm adjusts the recommendation strategy according to the learning state of learners, so as to provide personalized services for learners.

(2) *The basic principle of Ant Colony Algorithm:* Ant Colony Algorithm is a kind of simulated optimization algorithm which simulates the foraging behavior of ants. It was proposed by Italian scholar Dorigo M and others in 1991, and firstly used in solving TSP (Traveling Salesman Problem). The basic idea is:

- Ants release pheromones while foraging for food.
- They pick a road at random and release pheromones when they come across an intersection they haven't yet crossed. Pheromone concentration is inversely proportional to path length.

- Later ants choose the path with higher pheromone concentration when they encounter the intersection again.
- As the concentration of pheromones on the optimal path increases the colony finds the optimal feeding path.

The search process is listed as follows: at the initial time, the m ants are placed randomly in the city, and the initial pheromone values of each path are equal, set $\tau_{ij}(0)\tau_0$ as the initial pheromone value, $\tau_0 = m/L_m$, L_m is the path length. Next, according to the rule of random proportion, the ant K choose the next city to move, the selection probability formula is:

$$p_{ij}^{k}(t) = \begin{cases} \frac{\left[\tau_{ij}(t)\right]^{\alpha}\left[\eta ij(t)\right]^{\beta}}{\sum\limits_{s \in allowed} [\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}, j \in \text{allowed}_k \\ 0, \quad \text{otherwise} \end{cases} \tag{1}$$

in the formula, τ_{ij} is the pheromone on the edge (i, j), $\eta ij = 1/d_{ij}$ is heuristic factors for the transfer from city i to city j, allowed_k is the set of cities that ant K is allowed to visit in the next step, α is the information heuristic factor and β is the expectation heuristic factor.

In the process of pheromone updating, because the initial state is random, the volatilization factor ρ is often introduced in the Ant Colony Algorithm to eliminate the effect of the initial state. If the ρ is too small, a vast amount of pheromones will remain on each path, resulting in invalid paths continue to be searched and affecting the convergence rate of the algorithm. If the ρ is too large, the invalid path can be excluded, but it is possible that the efficient path may also be abandoned, thus affecting the optimal value of the search. Suppose that from time T to (T + 1) m ants are passing through the road i → j, the pheromone update formula is:

$$\tau_{ij}(t+1) = (1-\rho) \times \tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t) \tag{2}$$

in the formula, $\Delta\tau_{ij}^{k}(t)$ is pheromones accumulation that ants K retain on path i → j from time T to (T + 1).

However, the above algorithms still need to be improved:

- It is possible that the pheromone concentration on a certain path is much larger than that of other paths, which makes the algorithm converge to the local optimal solution prematurely.
- The algorithm may appear stagnation phenomenon, the search time is too long and the convergence speed is too slow.
- It is also found that the convergence of this algorithm is unstable through many experiments.

To solve the above problems, further research on Ant Colony Algorithm indicates that pheromone concentration depends on the value of volatile factor ρ, and pheromone concentration affects the convergence speed of the algorithm. Therefore, based on the learner group evaluation of learning path, learners' learning goal, knowledge level and learning style, this paper looks for the parameter value of volatilization factors. In the Python 3.7 environment, in the case of other parameters unchanged, the volatilization factors in the algorithm are changed, and the appropriate value of volatilization factors is found in many experiments to improve the stability of the algorithm.

3 Learning Path Recommendation Problem Description

3.1 Learner Characteristics

In the description of learning path recommendation, learners' characteristics mainly includes three aspects: basic information, learning style and knowledge level.

(1) *Basic information*: In the recommendation of learning path, the basic information of learners mainly includes learning goals.

(2) *Learning style*: according to the Felder-Silverman learning style model, the learning style is divided into four dimensions (information processing, perception, input, understanding), and the learning style vector is represented by S = (s_1, s_2, s_3, s_4). These four dimensions are further subdivided to divide the learning style into eight types (active/thought-provoking, perceptual/intuitive, visual/speech, sequential/comprehensive).

(3) *Knowledge level*: different learners have different knowledge degree of different subjects, such as in the postgraduate entrance examination review, the mathematics foundation of some students is better, then he is more relaxed in the mathematics subject review preparation, even in the learning path traversal process some knowledge points can be ignored directly; for cross-professional students may not have access to professional subjects knowledge, the most basic knowledge point should be considered in the learning path recommendation [10]. Therefore, the level of knowledge is also an important aspect of the complete description of learner characteristics, and further affects the learning path recommendation [11, 12]. Depending on the breadth and depth of relevant knowledge, the learning system can analyze the learner's knowledge level through simple tests and use L ($0 \leq L \leq 1$) for quantitative description. The closer the L is to 0, the closer the learner's knowledge level is to the beginner's level, and 1 means that the learner is close to the expert.

3.2 Learning Object Characteristics

(1) *Expression form of learning objects*: David A.Wiley consider learning objects to be any digital resource with reusable features and used to support learning. It has the characteristics of digitization, reuse and teaching oriented. Learning objects can use text, audio, video, image, interactive software and other ways to express knowledge. During the course of learning, learners usually review in combination with text, audio, video, image, interactive software and so on, e_i represents the proportion of knowledge content expressed in some form to the total knowledge content of the learning object, e_i need to meet the following conditions:

$$\sum_{i=1}^{n} e_i = 1(0 \le e_i \le 1) \tag{3}$$

(2) e_1, e_2, e_3, e_4 represent that a learning object is expressed by text, symbols (charts, animations, etc.), video (audio), interactive software, and vector $e = (e_1, e_2, e_3, e_4)$ is used to quantitatively describe the knowledge performance of the learning object, thus realizing the evaluation of the learning object.

(3) *Difficulty coefficient of learning objects*: in the process of learning object design on the learning path, the designer measures the difficulty degree by measuring the depth and breadth of the knowledge expressed by the learning object, and the difficulty coefficient values between [0,1]. the larger the value represents the more difficult to master the knowledge, the more suitable the learning object is for the expert.

3.3 Learning Path Recommendation

In real life, for learners who have already set their learning goals, the first problem to be solved is how to start the review of each subject. They will mainly refer to two aspects of information, one is the matching degree between learning style, knowledge level and learning object, the other is learning experience from learners who have achieved the same goal, to find a suitable path for their own review. But in the intelligent learning system, firstly according to the study subject it constructs the study object network diagram G. Then, according to the learning style and knowledge level, it recommend a group of learning objects in sequence $P = \{p_i, p_j,\ldots ,p_k\}$. After the learner completes the learning of all the learning objects in the learning path in turn, the learning object and the whole learning path are evaluated. After the evaluation information of the learning object is updated in the network diagram, it is used for the next learner's learning path recommendation.

(1) *Construction of learning object network graph*: using metadata, the system can establish a directed graph G (P, A) that describes the association relationship

of each learning object. In G, $O = \{o_1, o_2, \ldots, o_n\}$ represents a collection of learning objects that represent a particular learning task, $A = [\ a_{ij}]$ represents the adjacency matrix between learning objects, $0 \leq a_{ij} \leq 1$, $a_{ij} = 1$ means learning o_i before learning o_j, $a_{ij} = 0$ means that there is no connection between object o_i and o_j.

(2) *Learning path recommendation*: in order to help learners achieve the established learning goals, according to the characteristics of learners and learning object characteristics, using relevant algorithms to choose a suitable learning path for learners. Here, the learners' feature is the input of the algorithm and the learning path is the output of the algorithm. The selection of the algorithm needs the recommendation strategy based on the learning path, and the recommendation strategy needs to be analyzed in combination with the actual situation of the target user.

4 Learning Path Recommendation Based on Ant Colony Algorithm

Traditional Ant Colony Algorithm mainly has the following problems: artificial ants are not classified; there are two ways of local and global updating in the process of pheromone updating, which do not make full use of the optimal solution; the concentration of pheromone will affect the stability of Ant Colony Algorithm convergence, and the value of volatilization factor ρ associated with pheromone concentration is between [0,1], which lacks accuracy.

In this paper, learning style and knowledge level are used to identify similar learners. Global updating is used in the evaluation of learning path. Only pheromones on the optimal solution path are updated to achieve full utilization of the optimal solution. Selecting suitable volatilization factors ρ, improve the accuracy of the algorithm, thus enhancing the stability of the algorithm.

4.1 Similar Learner Recognition

The similarity is mainly manifested in three aspects according to the learners' characteristics: basic information, learning style and knowledge level. In the wisdom learning system, similar learners are identified in three steps. Firstly, the subjects are consistent, assuming that the learning group of a certain learning subject is $P_x = \{ p_1,p_2 \ldots p_n\}$; Secondly, in P_x, according to the Felder-Silverman learning style model, judging which learning style the learners belong to, $S_y = \{s_1,s_2 \ldots s_m\}$are learning groups with consistent subjects and same learning styles; Lastly, the knowledge level test is carried out in the S_y and the L $(0 \leq L \leq 1)$ is used to classify, so that similar learner group $Ls = \{l_1,l_2 \ldots l_k\}$ can be obtained.

4.2 Heuristic Information

In the intelligent learning system, the pheromone concentration is not the only factor considered by the learner to find the optimal path, but also the new section will be selected with a certain probability in the range of their own contact. In this paper, heuristic information refers to the learners after learning a learning object, will synthesize their learning style and learning object's expression form, their knowledge level and learning object's difficulty coefficient synthesis to make the choice.

(1) *Learning style*: in the previous learner's characteristic description and learning object's characteristic description, we can see that both learning style and learning object's expression form are vectors. In the intelligent learning system, the learning style of learner L_0 is set as s_0, the knowledge expression characteristics of learning objects o_j is set as e_j.The learning style heuristic information formula is:

$$\eta_{L_0 O_j}^{-S} = 1 - \left|s_0 - e_j\right| \tag{4}$$

(2) *Knowledge level:* set the difficulty of learning object o_j to d_j, the learner L_o knowledge level of the knowledge points attached to the o_j is l_0, then the knowledge level heuristic information formula is:

$$\eta_{L_0 O_j}^{-L} = 1 - \left|l_0 - d_j\right| \tag{5}$$

The heuristic information formula for integrating learning style and knowledge level is:

$$\eta_{L_0 O_j} = \left(1 - \left|s_0 - e_j\right|\right)^{\alpha 1} \times \left(1 - \left|l_0 - d_j\right|\right)^{\alpha 2} \tag{6}$$

α_1, α_2 are the information heuristic factors, the assumption in this article is $\alpha_1 = \alpha_2$.

4.3 Update on Pheromone

In the traditional Ant Colony Algorithm, the updating principle of pheromone is divided into two types: local update and global update. In order to make full use of the optimal solution and adopt the principle of global update in the intelligent learning system. After the learner completes the learning of all the learning objects in the learning path in turn, the learning object and the whole learning path are evaluated. After the evaluation information of the optimal path in the learning object network diagram is updated, it is used for the next learner's learning path recommendation. The pheromone intensity update formula is:

$$\tau_{ij}(t+1) = (1-\rho) \times \tau_{ij}(t) + \Delta\tau_{ij}^{best}(t) \quad \rho \in (0,1) \tag{7}$$

$$\Delta\tau_{ij}^{best}(t) = \begin{cases} \frac{1}{L^{best}} & \text{if } (i,j) \in T^{best} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

the specific values of the volatilization factor ρ will be discussed in the later simulation algorithm experiments.

4.4 Selection Probability of Learning Objects

Suppose a learner L_o completed the study of object o_i. According to the basic principles of Ant Colony Algorithm described above, the probability of selecting object o_j as the next learning object is:

$$p_{ij}^{L_0}(t) = \begin{cases} \dfrac{\left[\eta_{L_0 O_j}^{-S}\right]^{\alpha_1} \times \left[\eta_{L_0 O_j}^{-L}\right]^{\alpha_2} \times \left[\tau_{ij}(t)\right]^{\beta}}{\sum\left(\left[\eta_{L_0 O_j}^{-S}\right]^{\alpha_1} \times \left[\eta_{L_0 O_j}^{-L}\right]^{\alpha_2} \times \left[\tau_{ij}(t)\right]^{\beta}\right)} & \left(O_i \to O_j\right) \in J(i) \\ 0 & \left(O_i \to O_j\right) \notin J(i) \end{cases} \tag{9}$$

α_1, α_2 are the information heuristic factors, the assumption in this article is $\alpha_1 = \alpha_2$, β is the expectation heuristic factor. J(i) represents all possible learning paths starting from o_i.

4.5 Learning Path Recommendation Algorithm

- Leaner select target learning subject N.
- The learning object network diagram G is generated according to N, and section evaluation information is initialized.
- Building similar learners group L_s.
- According to the L_s learner's evaluation information of the path, the section evaluation information is updated.
- Creating ant colony $B = \{ b_1, b_2 \ldots b_n \}$.
- Ant Colony Algorithm initialization: learning path length and learning object.
- Cycle learning path length optimization calculation: For each bi in B, select the road section according to the probability selection formula, record path, revise of the Taboo Tab, updating road pheromones.
- Output the learning path according to the path evaluation value.

4.6 Simulation Algorithm Experiment

(1) *Experiment purpose*: on the basis of improving the traditional ant colony algorithm, the parameter volatilization factor ρ, which limits the pheromone concentration, is optimized. After a large number of experiments, the optimal value of volatilization factor is selected, so that the ant colony algorithm can be better applied to the learning path recommendation in intelligent learning.

(2) *Experiment design*: in the Python 3.7 experiment environment, the learning objects network diagram in the simulated intelligent learning system has 50 learning objects and 100 learners (assuming these are similar learners),the information heuristic factor $\alpha_1 = \alpha_2 = 1$ and the expectation heuristic factor $\beta = 2$. Since the initial learning object of different learner is different, the initial learning object is randomized in this experiment.
Three experiments were carried out in total, each experiment was divided into 9 groups: $\rho = 0.1$, $\rho = 0.2$, $\rho = 0.3$, $\rho = 0.4$, $\rho = 0.5$, $\rho = 0.6$, $\rho = 0.7$, $\rho = 0.8$, $\rho = 0.9$, each experiment runs 10 times, record the operation times that the algorithm gets the optimal solution. Then the average and standard deviation of each group are analyzed to find the value of volatilization factor with the best performance of the algorithm.

(3) *Experiment result analysis*: In order to make full use of the experimental data and ensure the accuracy of the experimental results, the experimental data will be analyzed from statistics and broken line graph respectively.

Calculate the average value and standard deviation for each group data. From the data in Table 1, it can be seen that the average of the algorithm operation times is 304 when $\rho = 0.5$. At this time, the algorithm runs the most efficiently.

From the broken line graph, it can be seen that the trend of the line is relatively stable when ρ = 0.5, which indicates that the standard deviation of the three experiments' results changes slightly and the algorithm is more stable (Fig. 1).

Table 1 Average value and standard deviation for each group data

ρ experiment	1	2	3	AVE
0.1	348	399	202	316
0.2	312	452	185	316
0.3	358	423	262	347
0.4	229	415	308	317
0.5	268	267	377	304
0.6	320	208	445	324
0.7	420	612	302	445
0.8	506	569	569	548
0.9	623	635	597	618

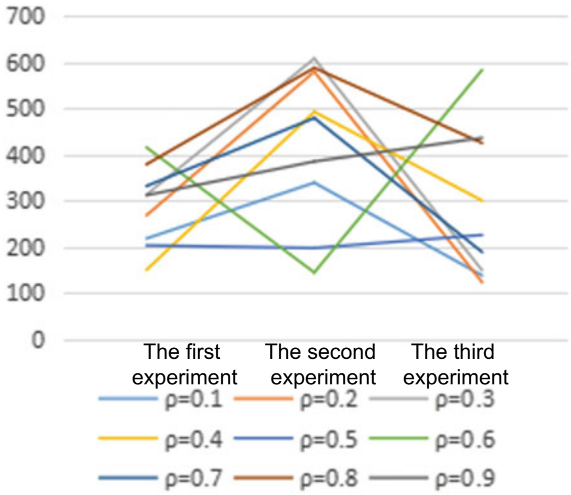


Fig. 1 Broken line graph of standard deviation of algorithm operation times

5 Conclusion

By adjusting the volatilization factor ρ in Ant Colony Algorithm, the learning path recommendation algorithm is optimized. And only learners' comments with similar learning style can be used as pheromone to guide recommendation decision. In the process of pheromone updating, learners' comments of learning objects on the optimal path and the comments of the whole optimal path are fully utilized by using the principle of global optimal updating. Experiment results show that when the volatilization factor $\rho = 0.5$, the extended ant colony algorithm has high efficiency and great stability. It can be applied to the learning path recommendation problem in intelligent learning, providing suitable learning path for learners, reducing blindness and randomness in learning, and improving learning efficiency.

According to the Research Triangle Model: generalizability, precision and realism, this paper adopts the experiment method. Although the precision of the method is high, it also has some limitations, such as the generalizability and realism is low, and there is a certain deviation between the experimental environment and the practical application.

References

1. B. Peter, Methods and techniques of adaptive hypermedia. User Model. User-Adapted Interact. **6**(2) (1996)
2. C.M. Chen, Intelligent web-based learning system with personalized learning path guidance. Chih-Ming Chen **51**(2) (2008)
3. T.C.R. Judy, H.C. Chu, H. Gwo-Jen, T. Chin-Chung, Development of an adaptive learning system with two sources of personalization information. Comput. Educ. **51**(2) (2007)
4. Y.C. Chang, W.Y. Kao, C.P. Chu, C.H. Chiu, A learning style classification mechanism for e-learning. Comput. Educ. **53**(2), 273–285 (2009)
5. Q. Zhao, A study on the recommendation of micro-learning path based on ant colony pheromone optimization algorithm. M.S. thesis, Taiyuan University of Technology ,Taiyuan, SX, China (2017)
6. B. Bert van den, T. Colin, J. José, B. Francis, K. Hub, K. Rob, Swarm-based sequencing recommendations in e-learning. Int. J. Comput. Sci. Appl. **3**(3) (2006)
7. Y.J. Yang, C. Wu, An attribute-based ant colony system for adaptive learning object recommendation. Expert Syst. Appl. **36**(2), 3034–3047 (2008)
8. Y. Cheng, Learning path recommendation based on swarm intelligence in online learning. J. Syst. Manage. **20**(2), 232–237 (2011)
9. Q. Zhao, J. Chen, Y.Q. Zhang, A recommendation of personalized micro-learning based on successive adaptive ant colony optimization algorithm. Comput. Eng. **44**(2), 238–243+276 (2018)
10. K. Moon, H. Kim, Performance of deep learning in prediction of stock market volatility. Econ. Comput. Econ. Cybern. Stud. Re. **53**(2), 77–92 (2019)
11. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik-Technical Gazette **25**(2), 528–535 (2018)
12. L.M. Wang, Z.Y. Hao, X.M. Han, R.H. Zhou, Gravity theory-based affinity propagation clustering algorithm and its applications. Tehnicki Vjesnik-Technical Gazette **25**(4), 1125–1135 (2018)

News Recommendation Based on Click-Through Rate Prediction Model

Guiying Wei, Yimeng Wei, and Jincheng Lei

Abstract News recommendation is one of the most popular applications in recommendation system, but the traditional recommendation algorithms are challenged by news features such as high update frequency, time-sensitive, high proportion of inactive users, large scale of news data, etc. In this paper we propose a news recommendation system based on click-through rate prediction model which has been used in online advertising, and data features are processed by one-hot encoding and gradient boosting decision tree. Comparative experiments proved its effectiveness and superiority in news recommendation.

Keywords News recommendation · Click-through rate prediction · Recommendation system · Gradient boosting decision tree · One-hot encoding

1 Introduction

With the development of information technology, we have entered the era of news explosion. Users access news more conveniently by all kinds of Internet platforms but they are also troubled by the exponential growth of news. How to present the news to users who are really interested in is an urgent problem to be solved.

Nowadays, news recommendation becomes a hot topic and research in this field can help users screen interested news easier. Specifically, collaborative filtering and content-based recommendations are the most popular recommendation algorithms. Collaborative filtering recommendation makes use of news ratings by users to provide recommendation services while content-based recommendation sequentially finds

G. Wei (✉) · Y. Wei · J. Lei
School of Economics and Management, University of Science and Technology
Beijing, Beijing, China
e-mail: weigy@manage.ustb.edu.cn

Y. Wei
e-mail: s20200975@xs.ustb.edu.cn

J. Lei
e-mail: leijincheng@ustb.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_27

newly-published news which is similar to the user's reading history in terms of content.

However, news recommendation has its different characteristics, such as high frequency of updating, time-sensitive, high proportion of inactive users, large scale of news data, etc. Since users' interest in reading news is often influenced by news trends, regional hot news and breaking news should also be recommended [1]. Therefore, traditional algorithms for news recommendation have their limitations and are not good at predicting user short-term news interest accurately. Additionally, the increasing complexity of recommendation is adverse to the iteration and updating of traditional recommendation models.

To solve above problems, extensive research work has been conducted on the news recommendation. In this paper, we propose a news recommendation system based on click-through rate (CRT) prediction model. Learned that the characteristics similarity between Internet advertising and news, we use the experience of estimating click-through rates of Internet advertising to forecast user's interest in the news, one-hot encoding and gradient boosting decision tree is used to preprocess the features. The proposed method can solve frequent cold-start problem and outperforms others in online updating of recommendation algorithms with gigantic data.

2 Literature Review

As an effective way and potential service to attract users, news recommendation systems have become popular in recent years. Personalized news recommendation has been extensively studied with the information overloading of news. Existing methods can be roughly classified into three categories: collaborative filtering, content-based filtering and hybrid recommendation algorithm [2]. Xia et al. proposed a hybrid news recommendation algorithm combining social relations and tag information [3]. Liu et al. utilized item-based incremental update collaborative filtering algorithm to adjust recommendation list dynamically [4]. However, in some scenario, simply traditional methods is insufficient to handle substantial amount of online users.

In view of the unique characteristics of news, a variety of studies have been proposed based on the difference between news and other item recommendations from multiple aspects. Personalized attributes of news recommendation data play a vital role in the study of this field. Joseph et al. utilized named entities to form a graph traversal model as well as a weighting scheme for cold-start content based news recommendation [5]. Categorize news topics with latent sematic features can promote exact news recommendation. Hu et al. constructed a graph neural news recommendation model with long-term and short-term interest model [6]. With pointed method, news can be recommended with higher accuracy.

Recently, deep learning solutions are attracted more attentions to improve personalized news recommendation tasks [7, 8]. Okura et al. utilized denoising autoencoder to analyze latent sematic information of news article [9]. Moreira et al. conducted a session-based news recommendation with recurrent neural networks which can leverage a variety of information types [10]. Achieving better performance in many cases, deep learning methods have drawn attention of some scholars.

The click-through rate prediction model was initially proposed in the Internet advertising and rapidly developed with the huge commercial interest [11]. Internet advertising has the characteristics of time-sensitive, higher frequency of updates and lower user's propensity to click on the advertisement, which are like the characteristics of online news. Click-through rate prediction model can be summarily categoried into two groups: traditional prediction model and deep learning model [12]. He et al. utilized the gradient boosting decision tree to extract features and predicted click-through rate by Stacking model [13]. Yang et al. further improved factorization machine supported neural network by adding a new feature generation layer and adopted inception structure for convolution [14]. Better performance can be expected by exploiting in-depth features and combining multiple methods.

Inspired by the advertising strategy of the internet advertising sector [15], in this paper, we propose a news recommendation system based on the click-rate prediction model. Utilizing the proposed model, we create user portraits based on the user behavior log to identify user interest and recommend news with high click-through rates by predicting whether users will click or be inclined to click on the displayed news. The novelty lies in the idea of incorporating click-through rate prediction model into the news recommendation. By experiment, we have proved the effectiveness of the proposed model.

3 Problem Definitions and Theoretical Principle

3.1 Problem Definition

In this paper, we use the users' click behavior to measure whether the user is interested in the news. Therefore, the news recommendation can be seen as a two-class classification problem: when the i piece of news is displayed, users either click on it or not, and $\mathrm{y_i}$ is used to indicate the classification result. If the user finally clicked on the news, $\mathrm{y_i} = 1$, otherwise $\mathrm{y_i} = 0$.

3.2 Theoretical Principle

3.2.1 Ranking Index

First of all, the recommendation system predicts the likelihood of users clicking on news. For news i, p_i denotes the probability of $y_i = 1$. For news i and j, if $p_i > p_j$, the news i will be ranked before the news j. In other words, news with high probability to be clicked has priority in user recommendation.

Confronted with ranking problem, we use the index area under curve (AUC) [16] to measure the accuracy of predicting the order of recommendation list. AUC is equivalent to the probability of $p_i > p_j$ given $y_i = 1$ and $y_j = 0$ ($0 \leq \text{AUC} \leq 1$). The closer AUC is to 1, the better the ranking is.

3.2.2 Numerical Index

For binomial problems, the logarithmic loss function can be used to evaluate the accuracy of the recommendation. We utilize Normalized Entropy (NE) [11] as the numerical index which denotes the ratio of the logarithmic loss to the total information entropy of clicking event. Noted that we utilize information entropy for normalization in order to offset the effect of large differences in the proportion of positive and negative samples in different datasets. The normalized entropy is shown in (1):

$$\mathrm{NE} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[-y_i \log p_i - (1 - y_i)\log(1 - p_i)\right]}{-p \log p - (1 - p)\log(1 - p)} \tag{1}$$

where p_i denotes the prediction probability of a user clicks on the news i and y_i represents whether news i will be clicked or not while p indicates the proportion of clicked news among n pieces of news. The smaller the normalized entropy is, the higher the accuracy of news recommendation is.

3.3 Click-Through Rate Prediction Model

Many characteristics of Internet advertising are similar to the news recommendations. Therefore, we study the characteristics of Internet advertising and click-through rate prediction model in advertising to solve the news recommendation problem.

Click-through rate prediction model was first proposed by Yahoo [17] and Google [18] in 2005–2007, it was used and promoted by Microsoft in its sponsor search engine advertising in 2007 [19]. Recently, most of existing studies in this field took advantage of machine learning approaches. Reference [20] proposed a novel framework for click prediction based on Recurrent Neural Networks.

In sponsor's search engine advertising system, when users send request to search engine, the advertising system will match the specific information with key words selected by advertisers for its advertising in advance. Then, the matched ads will be treated as the initial candidate ads set. At this time, in addition to the consideration of advertiser bidding factors, which ads are in the display candidate ad set and in what order are mainly determined by the probability user clicks on each ad when each ad is displayed based on the click-through rate prediction model [19].

In user set U, if user u sends request q, advertising system will match the candidate set A_q according to the request. For each ad a in the candidate set A_q, the main task of the click-through rate prediction model is to estimate the value of the function $P(\text{click} \mid q, a)$, which means the possibility that the user u clicks on the display advertisement a when searching for q.

The function $P(\text{click} \mid q, a)$ can be constructed by the logistic regression model which is represented by Google's Trenches system [19].

Click-through rate prediction model based on logistic regression utilizes logistic regression function to estimate the probability of clicking on a specific advertisement for a specific user. The main part of the logistic regression model is the sigmoid function which is extension and development of simple linear model. Logistic regression model is simple in form, easy to understand and widely used, and can learn latest data to update model, which is suitable for large-scale data processing. Since it has above advantages, we apply logistic regression model in our news recommendation system.

4 News Recommendation System

In this paper, we propose a news recommendation system as shown in Fig. 1.

To begin with, user requests for news recommendation to the front-end service. Secondly, the front-end service forwards the request to basic feature extraction service. Thirdly, the basic feature extraction service extracts basic features from four dimensions: scenes feature, media feature, user feature and news feature, and then forwards them to the feature processing service.

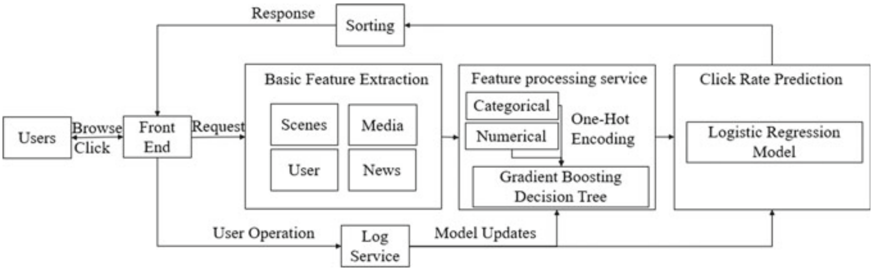


Fig. 1 Recommended system architecture

The feature processing service transfers the category features into numerical features by one-hot encoding, then inputs them into gradient boosting decision tree with original numeric features for automatic feature selection and combination, and forwards processed features to the click-through rate prediction model, which predicts the user's click-through rate for the news by logistic regression model based on the characteristics entered and transfers the result to the ranking service. The ranking service ranks candidate news and takes the top N news in front of the user.

Finally, the front-end service records the user's feedback behavior information (click or not) and sends it to corresponding log service for storage. These information can be used for regular offline updating of the decision tree model and online updating of the logistic regression model.

4.1 Basic Feature Extraction

The scene of news showing can be expressed as: in a scene, certain news is presented to users through certain media. Learned that, we explore basic characteristics from 4 dimensions: scene, media, users and news. Detailed information about can be seen in Table 1.

4.2 Feature Preprocessing

4.2.1 One-Hot Encoding

In the above model, basic features can be divided into two categories: categorical features and numerical features. In pre-processing step, we need to convert categorical features into numerical ones. At this point, we apply one-hot encoding which is simple and effective.

One-hot encoding can handle non-continuous news features without decoder and expand basic features to a certain extent. Assuming that there are n possible values of categorical features, they will be changed into n binary features by one-hot encoding. When certain categorical feature takes the i-th value, the i-th binary feature takes 1 and others take 0. For example, there are seven possible values for the week of the news show time, which corresponds to 7 binary features. When the news appears on Wednesday, the result of one-hot encoding is 0010000.

Table 1 Recommended the basic characteristics of the news

Dimensions	Feature	Feature Extraction Method	Examples
scene	Time	Through the server time to resolve the day of the week, specific time and time period	Friday, 8, morning
	Location	Through the visit ip analysis out of national, provincial and municipal	China, Yunnan Province, Kunming City
	Device	Through the user agent analysis out of operating system	Win7
	Explorer	Through the user agent analysis out of Explorer	IE8
Media	Web page features	Server Response Time (ms)	20
	News location	Where the news is displayed and the historical click rate for that location	2,0.003
Users	Demographics	Gender, age, age stage, occupation, position	Male, 27, middle-aged, teacher, Beijing
	Access habits	Access week distribution, access time period distribution, number of visits per week	(0.3, 0.2, …,0.1), (0.15, 0.1, …,0.03),4
	Long-term news interest	The click-to-click type distribution for past user-specific historical news over the past month	(0.2, 0.1, …,0.15)
	Short-term news interest	All-user history news click type distribution over the past hour	(0.34, 0.16, …,0.28)
News	Title	Clustering similar news into events and naming the event	Tiananmen snow carving in Changchun
	Release news	News release party, release time	China Tibet network,20,160,106 10
	Timeliness	Current time-Post time (hours)	2
	Types of news	The classification of news events	People's livelihood
	News summary	Using tf-idf method, extract the news content of key information	Tiananmen 0.3, snow sculpture 0.2, …

(continued)

Table 1 (continued)

Dimensions	Feature	Feature Extraction Method	Examples
	Historical CTR	Calculates the click-through rate for the past few hours of news	0.0025

4.2.2 Gradient Boosting Decision Tree

Since the click-through rate prediction model applies logistic regression model and the logistic regression model as a linear classifier is deficient in nonlinear classification ability, we construct the kernel function through the combination of basic features to make up it in our news recommendation system.

We use gradient boosting decision tree as the feature converter, the advantage is as followings: (a) It has the ability to automatically filter for important features. (b) Multiple nodes in the same decision tree form a combination of features. (c) It has strong generalization ability to promote method. In view of above points, the model can improve accuracy of the click-through rate prediction model and efficiency of news recommendation.

Gradient boosting decision tree [21] is a primary function which regards regression decision tree as addition model. It uses forward stagewise algorithm to make residual of previous model towards gradient descent direction and produces highly robust, interpretable procedures for regression and classification. The model is shown in (2):

$$\mathrm{f_M(x)} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \mathrm{T(x;\,\Theta_m)} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \sum_{\mathrm{j_m=1}}^{\mathrm{J_m}} \mathrm{c_{j_m} I\left(x \in R_{j_m}\right)} \tag{2}$$

where x denotes a feature vector, M represents the number of trees, $\mathrm{T(x;\,\Theta_m)}$ indicates regression decision tree, Θ_m indicates parameters of the tree, $\mathrm{J_m}$ denotes the number of leaf nodes in the m-th tree, and $\mathrm{c_{j_m}}$ represents the output constant value when x falls on the $\mathrm{j_m}$-th leaf node $\mathrm{R_{j_m}}$ of the m-th tree.

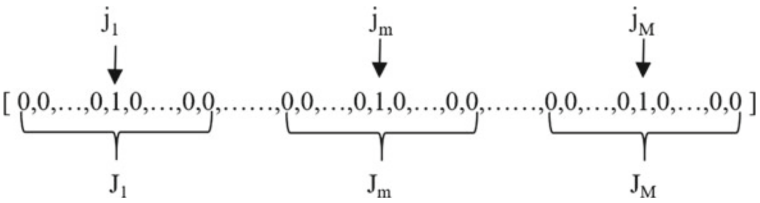


Fig. 2 The output of the feature conversion

The training process of gradient boosting decision tree is as follows:

```
input：$M$, $J=\{J_1,J_2,\ldots,J_m,\ldots,J_M\}$, $f_0(x)=0$
output：$f_M(x)$
for  $m=1$  to  $M$  do
    for  $i=1$  to  $n$  do
        $r_{m_i}=y_i\text{-}f_{m\text{-}1}(x_i)$
    end for
    fitting residuals  $r_m=\{r_{m_1},r_{m_2},\ldots,r_{m_i},\ldots,r_{m_n}\}$
    learning a regression decision tree and get  $T(x;\Theta_m)$
    $f_m(x)=\sum_{k=1}^{m} T(x;\Theta_k)$
end for
```

In the process of feature transformation, the model considers each decision tree as a categorical feature, and takes the index position of the leaf node where the sample data finally fall as the corresponding eigenvalue, then utilizes it for one-hot encoding. For example, the gradient boosting decision tree finally trains out M trees. When the sample data pass through the m-th tree, it falls into the j_m-th leaf node. Figure 2 shows the output of the feature conversion.

4.3 Click-Through Rate Prediction Model

The process of displaying n pieces of news is a Bernoulli process. The overall historical data and click data are consistent with a Bernoulli distribution. This is given by (3):

$$L=\prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{1-y_i} \tag{3}$$

where p_i denotes the prediction probability of the user clicking the i-th news and y_i indicates whether user click it.

Since (3) is a likelihood function that corresponds to a logarithmic loss function, it can be optimized by a logistic regression model. This is shown in (4):

$$h_w(x) = \frac{1}{1 + e^{-w^T x}} \tag{4}$$

where $p_i = h_w(x_i)$, and vector w is the weight of the feature and the best fitting parameter to be solved for the news click-through rate prediction model.

In order to solve parameter w, we have to apply optimization algorithm and choose Stochastic gradient descent algorithm [15] based on following two points: (a) It takes up less computational resources while the algorithm performs effectively. (b) As an online algorithm, it can update parameters when new data appear without the need to re-read entire dataset for batch processing. The core idea of Stochastic gradient descent algorithm is that each vector is updated with only one data. This is given by (5):

$$w_{new} = w_{old} - \eta\big(h_w(x_i) - y_i\big)x_i \tag{5}$$

where η indicates the learning rate.

In this paper, we improve the existing algorithm from two aspects. For one thing, adding norm to the logistic regression model to increase model parameters sparsity. By this way, model complexity can be reduced and generalization ability can be improved. The improvement makes the final training model smaller, easy to distribute to multiple servers for online click rate estimation and improve the response speed of the recommendation service.

Besides, utilizing the learning rate based on feature component to find the optimal fitting parameters w more efficiently. The eigenvector x is a high-dimensional sparse vector and the number of times that each feature component appears (the number of non-zero-valued data items for this feature component) varies greatly throughout the training dataset. Facing with above challenges, we apply learning rate to let each feature component use different updating steps, which makes it possible for the vector w to fully learn in different components and improve the accuracy of the click rate estimation model finally. This improvement is given by (6):

$$\eta_{t,i} = \frac{\alpha}{\beta + \sqrt{\sum_{s=1}^{t} g_{s,i}^2}} \tag{6}$$

where t represents the number of iterations of model updating, i denotes the index position in the d characteristic components, $g_{s,i}$ denotes the i-th component of gradient vector g_s in the iteration of the round s and $g_{s,i} = \big(h_w(x_s) - y_s\big)x_{s,i}$, α and β are tunable parameter ($0 \leq \alpha \leq 1$ and $\beta = 1$ under normal situations).

As a result, the process of online learning and forecasting for the click-through rate prediction model can be seen as follows:

```
input： $\alpha$, $\beta$, $\lambda_1$, $\lambda_2$, $\forall i \in \{1,\ldots,d\}, n_i=0, z_i=0$
output： $w=\{w_1,w_2,\ldots,w_d\}$
for t=1 to n do
    estimate the click through rate of t news, $p_t=h_w(x_t)=\frac{1}{1+e^{-w^T x_t}}$
    determine the value of $y_t$ based on user feedback
    for all $i \in \{i|x_{t,i} \neq 0\}$ do
        $g_i=(h_w(x_t)-y_t)x_{t,i}$
        $\sigma_i=\frac{1}{\alpha}\left(\sqrt{n_i+g_i^2}-\sqrt{n_i}\right)$
        $z_i \leftarrow z_i+g_i-\sigma_i w_i$
        $n_i \leftarrow n_i+g_i^2$
        $w_i=\begin{cases} 0, |z_i| \leq \lambda_i \\ -\left(\frac{\beta+\sqrt{n_i}}{\alpha}+\lambda_2\right)^{-1}\left(z_i-\mathrm{sgn}(z_i)\lambda_1\right), |z_i|>\lambda_i \end{cases}$
    end for
end for
```

5 Experimental Evaluations and Result Analysis

In order to evaluate the superiority of the proposed news recommendation method based on the click-through rate prediction model, we conduct the experiment by comparing different recommendation algorithms and analyze the proposed method from multiple index dimensions. All the algorithms are implemented by using Python and tested under the same experimental environment. In this paper, we apply the distributed cluster algorithm on Baidu cloud platform to handle large-scale data and the cluster nodes is about 500.

5.1 Dataset

The dataset is collected from the user behavior log data of news aggregating website. We sampled 1% traffic user behavior log data and related news date from April 19 to April 26, 2016. Note that the training set is collected from April 19th to April 25th and the test set is April 26th.

The original user behavior log data main record the information of user clicking history, including access to the proxy server, IP address, access time, and et al. Each user has a unique user ID for a long time tracking. Involved in daily data, the amount of users and log size approximately equal 100 thousand and 5 GB separately with news click rate around 1%. It shows that the amount of users and log size are large while the user behavior is relatively sparse.

5.2 Analysis of Results

In order to verify the effectiveness of our proposed model, we provide detailed comparison between ours and several mainstream algorithms in following experiments. We evaluate the AUC, NE, Precision, Recall, Coverage, information entropy where is denoted with H, Gini Coefficient where is denoted with G and Popularity based on the user behavior log data in the time range. The results are shown in Table 2.

As we have seen in Table 2, the content-based recommendation algorithm is the most effective one among the mainstream traditional recommendation algorithms. However, content-based recommendation algorithm can hardly solve the problem that high percentage of inactive users lead to data sparsity. In our dataset, 40% of users with click behavior on April 26 did not click from April 19 to April 25.

To mitigate the impact of data sparsity, we propose the click-through rate prediction model, the above experimental results show that the proposed model performs better in the evaluation of NE, AUC, Recall rate and Popularity. This can be attributed to three reasons: (a) The iteration of the proposed model is based on a logarithmic loss reduction. (b) We conduct our experiment in an incremental way. We add the user's characteristic and the context information to our model and train it to use more information from the dataset. (c) The recommendation algorithm based on the click-through rate prediction model is more time-sensitive than others, which lead to stronger ability in recommending novel news. Additionally, the proposed model performs more homogeneously according to the evaluation of H and G because it takes the content of the news into account and can mine long tail news to a certain extent.

Since the proposed model recommends news to users regardless of the historical click behavior, the evaluation of Precision and f1 value between the proposed model and content-based recommendation algorithm are equivalent while the proposed model achieve better results in terms of Coverage. Users without any historical click

Table 2 Results of collaborative filtering, content-based recommendation and click-through rate prediction

Algorithm	NE	AUC	Top	Accuracy (%)	Recall rate (%)	f1 (%)	Coverage (%)	H	G	Popularity
Item-based CF	326	0.51	5	7	12	9	26	0.42	0.94	6.1
			10	6	18	9	30	0.47	0.92	5.8
			20	5	23	9	35	0.51	0.88	5.5
User-based CF	364	0.51	5	6	5	6	23	0.43	0.94	6.1
			10	6	7	7	26	0.46	0.92	6.0
			20	6	9	7	27	0.49	0.90	5.8
Content-based	235	0.55	5	7	12	9	46	0.61	0.77	2.6
			10	7	21	10	48	0.62	0.74	2.6
			20	6	32	10	53	0.63	0.72	2.5
click-through rate prediction model	48	0.69	5	6	26	10	63	0.54	0.88	1.2
			10	6	40	10	81	0.59	0.82	1.4
			20	6	58	11	86	0.60	0.81	1.5

behavior may be inclined to click little news in the future. Therefore, the overall Precision and f1 value will be suppressed because of inactive users.

In summary, the recommendation algorithm based on the click-through rate prediction model can predict the users' behavior more accurately, which is reflected in covering more news and recommending the latest information among the mainstream recommendation algorithms included. On the other hand, it can solve cold-start problems to some extent, thereby enhancing the user experience and increasing user stickiness, and also handle the recommendation task of large-scale data as the iteration of logistic regression model updates online in real time.

6 Conclusions

In this paper, we proposed the news recommendation system based on click-through rate prediction model. The proposed algorithm can enhance the nonlinear classification ability of the linear regression model, solve frequent cold-start problem to a certain extent and handle online updating of recommendation models under large-scale data. Extensive evaluation has demonstrated the efficacy and efficiency of the news recommendation based on click-through rate prediction model.

For future work, to adjust the parameters and study how to improve them in a more quickly and effective way, such as the number of decision trees, the number of leaf nodes of each decision tree and the norm and learning rate of logistic regression model.

References

1. S. Wang, X. Li, F. Sun, C. Fang, Survey of research on personalized news recommendation techniques. J. Front. Comput. Sci. Technol. **14**(1), 18–29 (2020)
2. J. Liu, P. Dolan, E.R. Pedersen, Personalized news recommendation based on click behavior, in *Proceedings of the 15th International Conference on Intelligent User Interfaces* (ACM, 2010), pp. 31–40
3. H. Xia, C. Liu, Y. Liu, Hybrid news recommendation algorithm combining social relation and tag information. Appl. Res. Comput. **38**(1) (2020)
4. H. Liu, C. Wan, X. Wu, A hybrid recommendation model based on incremental collaborative filtering and latent semantic analysis. Comput. Eng. Sci. **41**(11), 2033–2039 (2019)
5. K. Joseph, H. Jiang, Content based news recommendation via shortest entity distance over knowledge graphs, in *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 690–699 (2019)
6. L. Hu, C. Li, C. Shi, C. Yang, C. Shao, Graph neural news recommendation with long-term and short-term interest modeling. Inf. Process. Manage. (2020)
7. G.D.S.P. Moreira, F. Ferreira, A.M.D. Cunha, News session-based recommendations using deep neural networks, in *The 3rd Workshop* (2018)
8. L. Huang, B. Jiang, S. Lv, Y. Liu, D. Li, Survey on deep learning based recommender systems. Chin. J. Comput. **41**(7), 1933–1942 (2018)

9. F. Wu, C. Wu, M. An, X. Xie, Personalized news recommendation based on deep learning. J. Nanjing Univ. Inf. Sci. Technol. (Nat. Sci. Ed.) **11**(3), 278–285 (2019)
10. G.D.S.P. Moreira, D. Jannach, A.M.D. Cunha, Contextual hybrid session-based news recommendation with recurrent neural networks. IEEE Access **7**, 169185–169203 (2019)
11. X. He, J. Pan, O. Jin, T. Xu, B. Liu, et al., Practical lessons from predicting clicks on ads at Facebook, in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising* (ACM, 2014), pp. 1–9
12. H. Liu, W. Yun, B. Lin, Y. Ding, A survey on feature learning and technologies of online advertising click-through rate estimation. J. Zhejiang Univ. (Sci. Ed.), **46**(5), 565–573 (2019)
13. X. He, W. Pan, H. Cheng, An advertisement click-through rate prediction model based on ensemble learning. Comput. Eng. Sci. **41**(12), 2278–2284 (2019)
14. Y. Yang, B. Han, Advertising click-through rate prediction model based on enhanced FNN. J. Nanjing Univ. Sci. Technol. **44**(1), 33–39 (2020)
15. H.B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, et al., Ad click prediction: a view from the trenches, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (ACM, 2013), pp. 1222–1230
16. T. Chen, L. Tang, Q. Liu, D. Yang, S. Xie et al., *Combining Factorization Model and Additive Forest for Collaborative Followee Recommendation* (KDD-Cup Workshop, Beijing, 2012)
17. B. Edelman, M. Ostrovsky, M. Schwarz, Internet advertising and the generalized second price auction: selling billions of dollars worth of keywords. Am. Econ. Rev. **97**(1), 242–259 (2007)
18. H.R. Varian, Position auctions. Int. J. Ind. Organ. **25**(6), 1163–1178 (2007)
19. M. Richardson, E. Dominowska, R. Ragno, Predicting clicks: estimating the click-through rate for new ads, in *Proceedings of the 16th international conference on World Wide Web* (ACM, 2007), pp. 521–530
20. Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, et al., Sequential click prediction for sponsored search with recurrent neural networks, in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, pp. 1369–1375 (2014)
21. J.H. Friedman, Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)

Data-Driven Phenetic Modeling of Scripts' Evolution

Gábor Hosszú

Abstract This paper presents an extended phenetic approach to classifying the examined historical scripts and determining some properties of their evolution. The main challenge in the phenetic modeling of historical scripts is the very large number of homoplasies, i.e. the coincidence of unrelated graphemes or writing rules in different scripts. A data-driven framework is proposed for evaluating the extended phenetic model of the examined scripts through the application of the parsimony principle of cladistics. The basic idea is to collect various evolutionary models for each grapheme and extend the phenetic model built on the matches these graphemes in various scripts. The combination of the phenetic model with particular evolutionary concepts of each grapheme results in an improved phenetic model, which is relatively protected from the effect of the homoplasies. To illustrate this framework, it was used to evaluate the extended phenetic model of four descendant scripts, including 117 features (characters in a cladistic sense).

Keywords Cladistics · Computational paleography · Data-driven evaluation · Group spectrum · Phenetics · Script spectrum · Scriptinformatics

1 Introduction

This paper presents an algorithm called *successive elimination* to evaluate the *extended phenetic model* of examined historical scripts. The main challenge in the phenetic modeling of historical scripts is the very large number of homoplasies, i.e. the existence of similar features (e.g. graphemes or writing rules) in various scripts without evolutionary relationship. In this paper, a data-driven framework for evaluating the extended phenetic model of the examined scripts are proposed. The basic idea is to collect various evolutionary models for each grapheme and extend the phenetic model built on the matches these graphemes in different scripts. The combination of the phenetic model with particular evolutionary concepts of each

G. Hosszú (✉)
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: hosszu.gabor@vik.bme.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_28

grapheme results in an improved phenetic model, which is relatively protected from the effect of homoplasies. The obtained extended phenetic model can then be evaluated based on the application of the parsimony principle of cladistics. To illustrate this framework, extended phenetic model of four descendant scripts, including 117 features (characters in a cladistic sense) is evaluated with the proposed successive elimination.

Scriptinformatics, as a branch of applied computer science, deals with discovering relationships between scripts and investigating the evolution of graphemes of scripts. In this sense, the script could be any cultural origin sequence of signs such as historical writing systems or urban graffiti. In scriptinformatics, the bioinformatics, the artificial intelligence, and machine learning tools are applied [1], including statistical analyzing the features (graphemes and writing rules) of the scripts [2, 3]. A goal of scriptinformatics is to determine instances of genetic relatedness amongst scripts. *Computational paleography* is a special field of the scriptinformatics, which is focusing on the evolution of historical scripts, processes data of the written cultural heritage. It provides support for deciphering ancient inscriptions [4, 5]. Computational paleography, and in general, scriptinformatics is similar to computational finance [6] in that sense that the latter also uses artificial intelligence methods to describe a certain field of science.

The *script* is the graphic representation of a *writing system*, which is "a set of visible or tactile signs used to represent units of language in a systematic way" [7]. The script includes both grapheme set and writing rules. The *inscription* is a script relic of one or more scripts independently of the writing materials (ink and paper/papyrus/parchment, stone, wall, wood, etc.), and physically it can be a codex, an epigraph, a fragment, a manuscript, or a scroll. The *grapheme* ([in the computer science sense] character) is the minimum semantically or phonetically distinguishing component of a script [8]. The *glyph* is a grapheme pattern that appears on inscriptions, often with various topological parameters.

The research on scriptinformatics examines evolutionary changes in scripts through the development of graphemes and writing rules. To describe the changes of graphemes, a multilayer grapheme model was created composed of logical layers from top to bottom, namely the Semantic, Phenetic, Visual Identity, and Topology Layers. The Semantic Layer determines the context of the grapheme in the inscriptions, the Phenetic Layer provides its sound values, the Visual Identity Layer focuses on the unique identity of a grapheme based on the human visual perspective in identifying an object, and the Topology Layer describes a glyph by geometrical parameters [9].

To analyze research trends and scientific production, the scientometric examination carried out based on the citations of scientific publications [10]. In scriptinformatics, an ontology is defined by determining categories of evolutionary computing scripts [11]. The ontology of scriptinformatics encompasses the approaches to model the properties of scripts. Namely, the main elements of this ontology are the graphemes with its properties (glyphs, use periods, sound values), the taxon-feature scenario (the taxa are the scripts, and the features are the glyphs of graphemes belonging to the particular scripts), and the applied phenetic or cladistic methods [1].

In knowledge engineering, there are data-driven solutions [12] for the problem of how to measure the appropriateness of the existing ontologies for a certain field.

Belmonte et al. performed a statistical analysis of grapheme sets of various Libyco-Berber inscriptions [13]. (Libyco-Berber script was in use from 9th/7th c. BC to AD 7th c. in the Canary Islands and Northwest Africa.) Their performed analysis resulted in a dendrogram of the similarity of the grapheme set of the different inscriptions. Finally, they derived some consequences from their results to the geographic origin of the Libyco-Berber script.

Skelton applied two methods of the phylogenetic systematics for reconstructing the Ancient Greek dialects; namely, the feature [in that paper "character"] weighting and preliminary cluster analysis [14]. Revesz used bioinformatic evolutionary tree algorithms and applied similarity measure between pairs of graphemes of scripts [15]. Moreover, Revesz et al. developed the AIDA, the Ancient Inscription Database and Analytics system to aid researchers in studying various historical scripts, survived relics, and their possible readings [16].

Hosszú applied phylogenetic methods to model the evolution of historical scripts [1]. Phylogenetics aims to explore the evolutionary relationships between taxa to get an understanding of their evolution. In biological evolution and generally in phylogenetics, the taxa (taxonomic units) are usually species. Analogously in scriptinformatics, the script is usually the taxon. In biological evolution, the taxon (species) consists of living organisms. By analogy, a historical script (taxon) in the scriptinformatics consists of surviving inscriptions written in that script. In phylogenetics, a feature (the character in a cladistic sense) is a heritable trait of a taxon. It is noteworthy that in pattern recognition, the term "character" is used similarly to the term "grapheme". In scriptinformatics, the terminologies of both pattern recognition and phylogenetics are applied, which makes the term "character" ambiguous. That is why in scriptinformatics instead of "character", the preferred terms are "feature" and "grapheme," respectively [1]. The *feature* in scriptinformatics can be a grapheme or a writing rule. E.g., the two *states of a feature* (if the feature is a grapheme) are "the presence of a glyph variant of a grapheme" and "absence of that glyph variant."

Phenetics (numerical taxonomy) and cladistics (phylogenetic systematics) are two areas of phylogenetics. Phenetics uses phenotypic data, and cladistics explores genealogical information. In cladistics, *apomorphy* is a derived feature state of a taxon, and *homoplasy* is when two or more apomorphic feature states are identical; although they originated from not a common ancestor, but rather by convergence or reversal. Because of the scarcity of paleographic information in case of several historical scripts, there usually is not enough information to model the phylogenetic relationships of scripts; therefore, basically phenetic analysis is performed to reveal the similarity relationships between scripts. Hosszú developed a *phenetic model* of 66 different scripts with 186 different features [1]. The specialty of the scripts that there are several homoplasies due to the limited topological variations of the simplest glyphs. Therefore, the pure phenetic approach to describe the similarity of scripts can lead to erroneous results due to incorrectly linked glyph variants of genealogically unrelated scripts. In this paper, an extended phenetic approach is presented to make the phenetic model less vulnerable to homoplasies.

This paper is organized as follows: Sect. 2 presents the extension of the phenetic method, Sect. 3 presents the evaluation of the extended phenetic model, and Sect. 4 concludes.

2 Extended Phenetic Model

2.1 The Problem of the Homoplasies

In phenetic modeling, a two-dimensional binary taxon-feature data matrix can be created based on the presence or absence of features (glyph, grapheme, and writing rule) in scripts (taxa) [1]. In this *simple phenetic model,* there is no distinction between ancestor and descendant scripts in case of each feature. Therefore, the scripts having a certain feature creates a specific *Similarity Features Group* (SFG). This is as if all the scripts in the SFGs were a descendant script. The price of this unambiguousity was that each SFG contained all the scripts that contain a particular feature (grapheme or writing rule). In this way, homoplasies were not filtered out that is, in the various scripts, seemingly identical but evolutionary independent features were included in the same SFG. This reduced the accuracy of the simple phenetic model.

In *feature engineering,* it is necessary to decide for each taxon examined whether or not each feature is present. However, in case of the scriptinformatics, it is often not possible to determine whether the similarities between the two glyphs are evolutionary or merely incidental (homoplasy). In this case, the decision can lead to a mistake that degrades the goodness of the phenetic model. There was this kind of uncertainty about the simple phenetic model of the historical scripts [1]. Earlier authors (e.g., [17, 18]) tried to explore the evolutionary relationships between graphemes (comparative paleography); however, they cannot handle the inherent uncertainties of the glyphs' development.

To overcome the limitations of the previous simple phenetic model [1], evolutionary considerations are included in the analysis of a feature for a studied, descendant script (taxon), which is a feature of cladistics rather than phenetics. The model created in such a way called *extended phenetic model* (P_e). It should be noted that evolutionary concepts (origin models) introduced into the extended phenetic model are specific to a particular feature, and do not necessarily extend to other features of the descendant script. There are many features where only one origin model can be considered. Still, most of the features have more than one alternative, and it is not possible to determine with certainty which description of the evolutionary step that occurred. In other words, based on the available data, it is often not possible to determine which ancestor script of a given descendant feature comes from. Because there may be several alternatives to a descendant feature, they are collectively included to the same SFG. For each feature (grapheme or writing rule) of the descendant scripts an SFG was created, which includes the possible ancestor features of scripts assumed to be an ancestor. If in an SFG there are more than one ancestor script, only one of

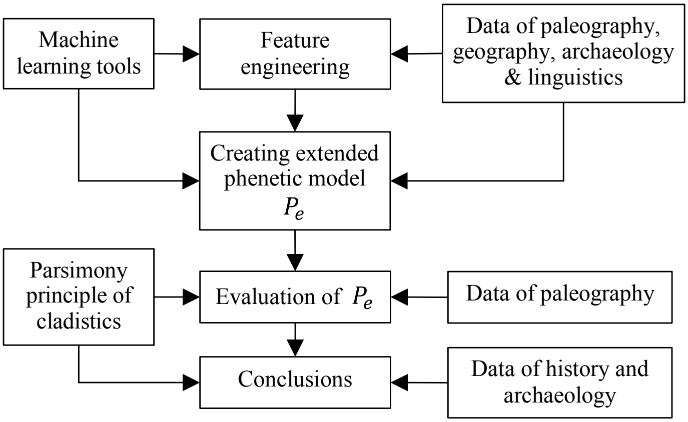

Machine learning tools
Feature engineering
Data of paleography, geography, archaeology & linguistics
Creating extended phenetic model P_e
Parsimony principle of cladistics
Evaluation of P_e
Data of paleography
Conclusions
Data of history and archaeology

Fig. 1 The flow of the extended phenetic analysis

them is the real ancestor with respect to that SFG. Despite this, in the following, any script (except the studied, descendant scripts) is called *ancestor script* if there is at least one SFG, in which there is an ancestor feature belongs to this script. Hence the name "ancestor script" is only valid in the context of a particular phenetic model.

The extended phenetic model P_e, which is based on various evolutionary considerations (origin models) for describing the studied, descendant scripts, is ambiguous since some of the SFGs that make up the origin model contain ancestor features belonging to more than one ancestor scripts [3]. One reason for this is that SFGs have multiple alternative origin models that describe different evolutionary backgrounds (geographic, historical, linguistic data) and that the origin models in the same SFG usually contain ancestor features belonging different ancestor scripts. Another reason for the diversity of ancestor scripts in some SFGs is that even a unique origin model often results in more than one possible ancestor feature (grapheme or writing rule). Namely—e.g. in case of graphemes—some graphemes in different scripts can be very similar; therefore, based on the shapes and sound properties, it is often not possible to decide which ancestor script a descendant grapheme comes from.

Due to its ambiguity, a simple phenetic model (two-dimensional binary taxon-feature data matrix) has to be created from the extended phenetic model P_e for a phenetic analysis. This step is called evaluation of P_e, and its will be details in the following. The flow chart of the elaborated extended phenetic analysis is presented in Fig. 1.

2.2 Descendant and Ancestor Scripts

In the extended phenetic analysis, the scripts are distinguished as "descendants," "ancestors," and others. All features of the descendant scripts have been included in the extended phenetic model, whereas only that features of the ancestor scripts are in the model, which match the appropriate features of the studied, descendant scripts. Therefore, the extended phenetic model is comprehensive for the descendant scripts, only.

In the present research, the studied, descendant scripts are the Rovash scripts used by certain people in the Eurasian Steppe; namely, the Turkic Rovash (Turkic runic, TR), Székely-Hungarian Rovash (SHR), Carpathian Basin Rovash (CBR), and Steppe (Steppean) Rovash (SR) [4]. Of these four scripts, only the SHR remained in use until today, while the other three become extinct after the eleventh century AD [19]. The common term "Rovash" comes from the Hungarian name of these scripts: *rovásírások* 'rovash scripts'.

The TR, and thus probably all Rovash scripts, originate in the region of the Inner Asian Altai Mountains [20]. However, around 600 BC, a presumably Cimmerian population migrated from the region of Gordion in Asia Minor (Anatolia) to the Altai [21]. In principle, this population could bring literacy with them, although there is no evidence to that effect. Therefore, we must include in our study the scripts used in the seventh century BC in Anatolia as potential ancestors.

While in the present phenetic analysis there are four descendant scripts, the number of potential ancestor scripts is 26. Many of them are related to each other, and their features in some cases are barely distinguishable. It is, therefore, advisable to treat them as script groups. Furthermore, scripts used at the same period, in the same region, but an evolutionary distance from one another, have been merged into a group for easier of use (Table 1).

2.3 Processing Operators

It is not possible to determine whether the existing features in two scripts are evolutionary related or their similarity is a mere coincidence, even in the extended phenetic model P_e. However, while constructing a simple phenetic model had to make a probabilistic decision for each feature during the feature engineering phase, for building the extended phenetic model P_e, this decision about homoplasy is bypassed, and all reasonable options are incorporated into the alternative origin models in each SFG. Moreover, alternative ancestor scripts may be included within each origin model in each SFG if necessary. An origin model in an SFG generally refers to a group of ancestor scripts, and it is often not possible to decide, which is the actual ancestor, so there may be several potential ancestor scripts in an origin model of an SFG.

The consequence of the above is that most of the SFGs in the extended phenetic model P_e are ambiguous in that sense that each of their descendant features have more than one ancestor feature in the same SFG. Thus, these SFGs should be excluded from further analysis during the evaluation of P_e. However, this would result in a significant loss of information. In the present state of the research, the extended phenetic model P_e contains 117 SFGs, of which only 51 SFGs have an ancestor feature of a unique

Table 1 Groups of potential ancestor scripts

Script group	Component script	Possible impact area on Rovash scripts
Aegean, Canaanite, Anatolian Syllabic and Anatolian-Greek Alphabetic	Cypro-Greek, Phoenician, Old Aramaic, Anatolian Hieroglyphic, Old Phrygian, Ancient Greek, Lydian and Carian	Anatolia
Aramaic and Late Aramaic	Imperial Aramaic, Syriac and Armazian	Inner Asia
Greek Alphabetic	Greek and Greco-Bactrian	Inner-Asia, Pontus Steppe or Carpathian Basin
Middle Iranian	Parthian, Khwarazmian, Sogdian, Middle Persian, Manichean, Avestan and Christian Sogdian	Inner-Asia
Indic	Kharoshthi, Brahmi and Tibetan	Inner-Asia
Slavic	Glagolitic and Early Cyrillic	Carpathian Basin
Latin Alphabetic	Latin	Carpathian Basin
Rovash (internal development)	Turkic Rovash (TR), Székely-Hungarian Rovash (SHR), Carpathian Basin Rovash (CBR) and Steppe Rovash (SR)	Inner-Asia, Pontus Steppe or Carpathian Basin

ancestor script, and the other 66 SFGs have more potential ancestor features from multiple ancestor scripts.

There are SFGs, which contain descendant features result of purely internal development, i.e., in the present research, the ancestor features in that SFGs belong to the Rovash scripts (Table 1). The Rovash scripts had presumably a common ancestor (called Proto-Rovash [1]). If an SFG has more than one ancestor feature belonging to multiple Rovash scripts, the real ancestor is probably the presumed common progenitor of the Rovash scripts, the supposed Proto-Rovash. Hence, the SFGs with purely internal development is always considered to have only one ancestor script and counted in the unambiguous SFGs that have only one, unique ancestor script.

To evaluate the extended phenetic model P_e, it is necessary to make P_e unambiguous even at the cost of reducing its accuracy in some way. The extended phenetic model P_e thus transformed into a so-called *simplified phenetic model* [3], which is a kind of simple phenetic model. This transformation step can be understood as a *processing operator*. A processing operator performs a transformation on the phenetic model as an argument and creates a more explicit, simplified phenetic model. In the following, some processing operator will be introduced. The series of the processing operators together create a data-driven algorithm called successive elimination.

3 The Successive Elimination and a Case Study

3.1 Generating Script Spectra and Group Spectra

In the following, the successive elimination algorithm is presented with a case study. The SFGs with multiple ancestor graphemes are ambiguous concerning ancestor scripts. The simplest evaluation of the extended phenetic model P_e is to consider only those SFGs that contain a unique ancestor script. Let E_s be an evaluation processing operator that generates a simplified phenetic model containing SFGs with unique ancestor script in each SFG. If the argument of the operator E_s is P_e, then the result is a simplified phenetic model of P_{oneS}, see (1).

$$P_{oneS} = E_s P_e \tag{1}$$

The so-called *script spectrum* of each descendant script can be generated based on the number of SFGs in P_{oneS}, see Fig. 2. It is noteworthy that these script spectra were generated based on only 51 SFGs of the total 117 SFGs, since the operator E_s skips the SFGs with multiple ancestors when creating P_{oneS} (Table 2). However, if the ancestor scripts of an SFG belong to the same script group (Table 1), then the uncertainty can be eliminated by specifying only the ancestor script group instead of the ancestor scripts since these SFGs have only one ancestor script group; therefore, these SFGs become unambiguous. This increases the number of unambiguous SFGs, and thus the reliability of the evaluation. By using the script groups, the extended phenetic model P_e can be evaluated by considering the SFGs with a unique ancestor script group rather than a unique ancestor script applying the processing operator

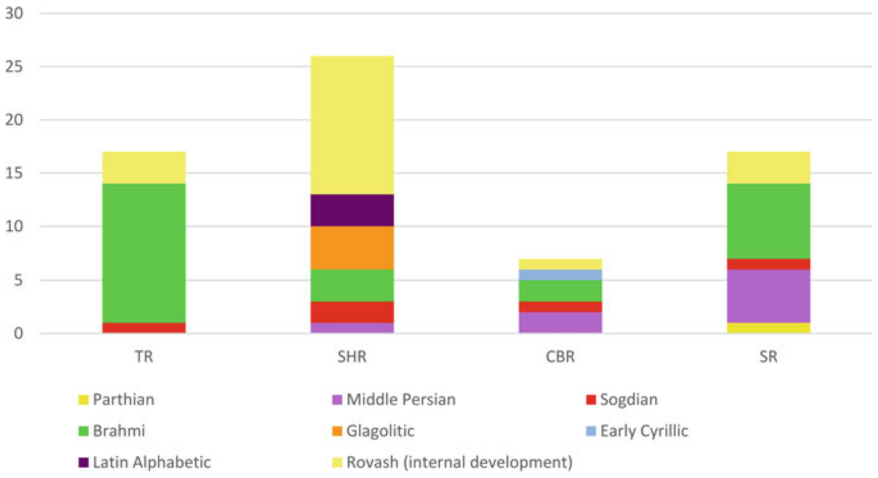


Fig. 2 Script spectra of each descendant script based on the simplified phenetic model P_{oneS}

Table 2 Comparing unambiguous phenetic models

Simplified phenetic model	Number of SFGs	Percentage of the number of SFGs in $P_{E\ (\%)}$
P_{oneS}	51	44
P_{oneG}	57	49
$P_{oneSAncG}$	75	64
$P_{oneGAncG}$	93	79
$P_{oneSAncS}$	82	70
$P_{oneGAncS}$	95	81

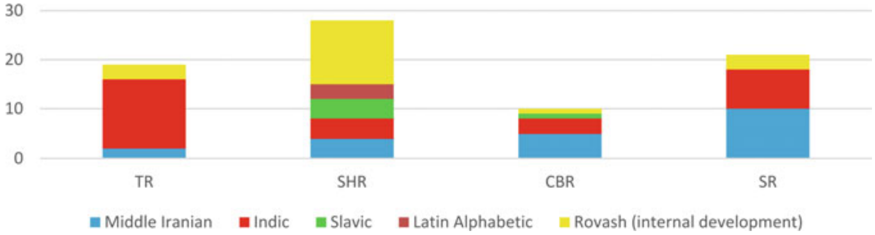


Fig. 3 Group spectra of each descendant script based on the simplified phenetic model P_{oneG}

called E_g. The result of E_g is P_{oneG} according to (2). P_{oneG} is based on 57 SFGs with unique ancestor script groups (Table 2). *Script-level unambiguous SFG* means an SFG with unique ancestor script, *group-level unambiguous SFG* is an SFG with unique ancestor script group.

$$P_{oneG} = E_g P_e \tag{2}$$

The *group spectrum* of each descendant script group can be created based on the number of SFGs in P_{oneG}, see Fig. 3.

3.2 Eliminating the Less Likely Script Groups

The evaluation of the extended phenetic model can be further improved by taking into account that a number of the ancestor scripts are not unique ancestors in any SFG. Moreover, the groups of some of these ancestor scripts are not unique ancestor script groups in any SFG. From the literature on the evolution of scripts, it seems that the various historical scripts adopted graphemes from a relatively small number of other scripts. Therefore, probably a correct evaluation strategy is that the ancestor scripts being not a member of a group, which is a unique ancestor script group at least in one SFG, these scripts are unlikely to be the ancestor of the examined descendant scripts.

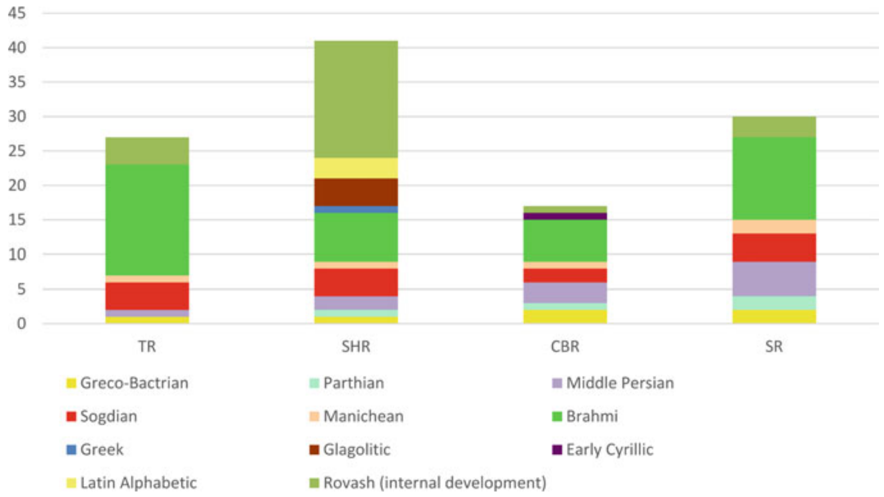


Fig. 4 Script spectra of each descendant script based on the simplified phenetic model $P_{oneSAncG}$

This strategy is implemented by the E_{ancG} processing operator, which transform the original P_e model to the simplified phenetic model P_{ancG} according to (3) by omitting that ancestor scripts from the SFGs, which are not unique ancestor script in any SFG and not a member of a group being unique ancestor group at least in one SFG.

$$P_{ancG} = E_{ancG} P_e \tag{3}$$

The number of SFGs in P_{ancG} equals to the number of SFGs in P_E, 117. It is noteworthy that script spectrum or group spectrum can be generated from an only unambiguous phenetic model, i.e., such model contains only script-level unambiguous SFGs or (at least) only group-level unambiguous SFGs. Such unambiguous models can be generated by the processing operators E_s and E_g. Generally, the processing operator E_{ancG} does not create an unambiguous simplified phenetic model. In the evaluation of the ambiguous simplified phenetic model P_{ancG} the already introduced E_s and E_g processing operators should be applied to it to obtain an unambiguous model. By using E_s, the script-level unambiguous SFGs of $P_{oneSAncG}$ are selected (4), and by applying E_g, the group-level unambiguous SFGs of $P_{oneGAncG}$ are collected (5).

$$P_{oneSAncG} = E_s P_{ancG} \tag{4}$$

$$P_{oneGAncG} = E_g P_{ancG} \tag{5}$$

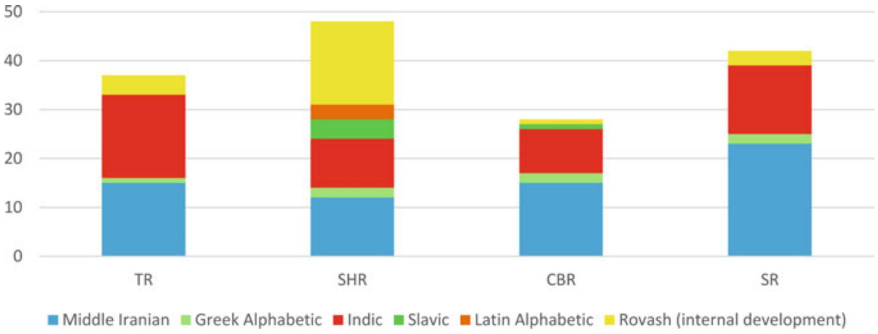


Fig. 5 Group spectra of each descendant script based on the simplified phenetic model $P_{oneGAncG}$

The script spectra of the descendant scripts (Fig. 4) is generated based on the number of SFGs in $P_{oneSAncG}$, which is 75, a significant increase over the number of SFGs in P_{oneS}, see Table 2.

The group spectrum of each descendant script group (Fig. 5) is created based on the number of SFGs in $P_{oneGAncG}$, which is 93, a dramatic increase over the number of SFGs in P_{oneG}, see Table 2.

Note that in the procedure E_{ancG} sort order is needed to set up when examining ancestor script groups. First, the earliest ancestor script group that is not unique to any SFG should be examined. If this is omitted, some of the remaining ancestor script groups may become unique in some SFGs. Then, of the remaining ancestor script groups that are not unique to any SFG, the earliest should be examined again. This process takes place until there are no ancestor script groups that are not unique to any of the SFGs. E.g., the Greek Alphabetic ancestor script group was not originally unique in any SFG (that is why it does not appear in Fig. 3); however, by omitting the Anatolian scripts, and the Aramaic and Late Aramaic scripts, the Greek Alphabetic became unique in some SFGs. That is why the Greek Alphabetic presents in Fig. 5. Without sorting, the Greek Alphabetic ancestor script group would have dropped out in the procedure E_{ancG}.

3.3 *Eliminating the Less Likely Scripts*

In the P_{ancG} simplified phenetic model (3), there are ancestor scripts, of which script groups appear as the unique ancestor script group in at least one SFG, but not as a unique ancestor script. Therefore, it can be assumed that those members of an ancestor script group, which never appear as unique ancestor script may not have influenced the descendant scripts, merely their similarity to the true ancestor scripts resulted in their being in a unique set of ancestor script groups. Based on this, the E_{ancS} processing operator creates the P_{ancS} simplified phenetic model (6) by omitting the ancestor features from the SFGs of the simplified phenetic model P_{ancG} that are not unique ancestor scripts in any SFG.

$$P_{ancS} = E_{ancS} P_{ancG} \tag{6}$$

The number of SFGs in P_{ancS} equals to the number of SFGs in P_{ancG}. In the evaluation of P_{ancS} the E_s and E_g processing operators can be applied to it. By applying E_s, the script-level unambiguous SFGs of $P_{oneSAncS}$ are gathered (7), and by applying E_g, the group-level unambiguous SFGs of $P_{oneGAncS}$ are chosen (8).

$$P_{oneSAncS} = E_s P_{ancS} \tag{7}$$

$$P_{oneGAncS} = E_g P_{ancS} \tag{8}$$

The script spectrum of each descendant script (Fig. 6) can be generated based on the number of SFGs in $P_{oneSAncS}$, which is 82, a significant increase over the number of SFGs in $P_{oneSAncG}$, see Table 2.

The group spectrum of each descendant script group (Fig. 7) can be created based on the number of SFGs in $P_{oneGAncS}$, which is 95, a little increase over the number of SFGs in $P_{oneGAncG}$, see Table 2.

The comparison of the number of SFGs in the unambiguous simplified phenetic models used for creating script and group spectra is presented in Table 2.

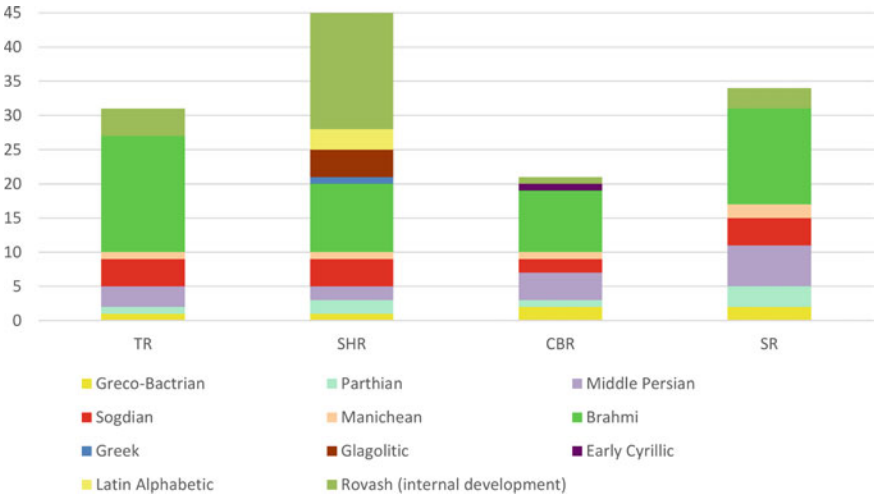


Fig. 6 Script spectra of each descendant script based on the simplified phenetic model $P_{oneSAncS}$

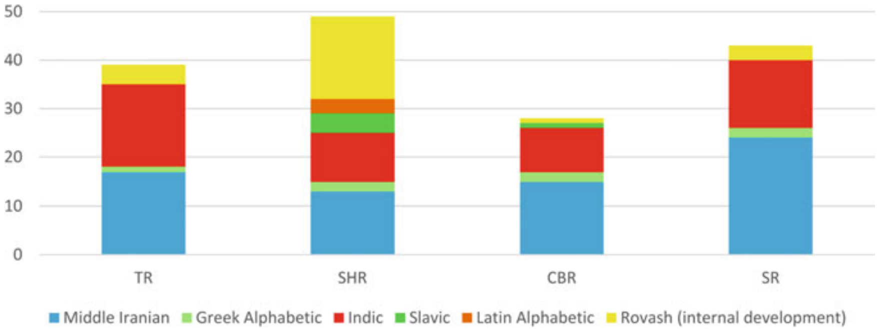


Fig. 7 Group spectra of each descendant script based on the simplified phenetic model $P_{oneGAncS}$

4 Conclusions

The paper described an extension to phenetic analysis that takes into account the peculiarity of the script evolution that there is a large number of similar, but evolutionary independent features (graphemes or writing rules) in scripts. Although the phenetic model does not usually contain evolutionary information, the extended phenetic model is designed to take into account evolutionary data as well as similarity data.

An evaluation algorithm called successive elimination was also detailed. From the resulting script and groups spectra, some conclusions can be drawn about the evolution off the examined—Rovash—scripts. First of all, while theoretically possible for these scripts to originate from Anatolia, it seems from the successive elimination of the extended phenetic model that the assumption of an Anatolian origin is not necessary to create an evolutionary model. Moreover, even a relationship with the very early Imperial Aramaic script is not justified. Taking into account the parsimony principle of cladistics [22], these scripts can be omitted from the ancestors. Instead, the two most important roots of Rovash scripts are the Middle Iranian (mainly Sogdian) and the Brahmi from the Indic scripts. The results obtained demonstrate that the Székely-Hungarian Rovash, the Carpathian Basin Rovash and the Steppe Rovash scripts were less influenced by the Brahmi script than the Turkic Rovash. It follows that the users of the previous three Rovash scripts left Inner Asia while the influence of the Brahmi script on Rovash scripts was still in process, and henceforth it could only affect the Turkic Rovash. The influence of the Greek, Glagolitic, Early Cyrillic and Latin scripts were able to influence Rovash scripts only in the Carpathian Basin, so their effects could not be felt until after the Rovash scripts split. The results obtained are in good agreement with Babayarov' statement that the Old Turkic coins had legend written in Sogdian script mostly; however, there are legends in Middle Persian, Bactrian, Ancient Indian and Turkic Rovash scripts, as well [23].

The presented results proved that phenetic modeling extended with evolutionary considerations are suitable for describing the similarities of historical scripts, and can even draw evolutionary conclusions. The data-driven evaluation method of the extended phenetic model called successive elimination was also presented. It has been demonstrated that the conclusions drawn are compatible with the result of history and humanities-like paleography. Thus, the phenetic modeling can become a useful tool for understanding the evolution of scripts.

References

1. G. Hosszú, Phenetic approach to script evolution, in *Kodikologie und Paläographie im Digitalen Zeitalter 4—Codicology and Palaeography in the Digital Age 4*, ed. by H. Busch, F. Fischer, P. Sahle (Books on Demand, Norderstedt, 2017), pp. 179–252
2. G. Hosszú, Mathematical statistical examinations on script relics, in *Data Mining and Analysis in the Engineering Field*, ed. by V. Bhatnagar (Information Science Reference, Hershey, PA, 2014), pp. 142–158
3. G. Hosszú, Írásemlékek grafémaalakjainak térstatisztikai és fenetikai elemzése [The Spatial Statistical and Phenetic Analysis of the Glyphs of Script Relics] *in Rovás – magyar nyelvtörténet – művelődéstörténet*, ed. by Zelliger (Magyarságkutató Intézet, Budapest, 2019), pp. 120–450
4. G. Hosszú, *Heritage of Scribes. The Relation of Rovas Scripts to Eurasian Writing Systems*, 3rd edn. (Rovas Foundation, Budapest). Available: https://google.hu/books?id=TyK8azCqC34C&pg
5. L.L. Tóth, G. Hosszú, A new topological method for examining historical inscriptions. J. Inf. Technol. Res. (JITR) **12**(2), 1–16 (2019)
6. E.P.K. Tsang, S. Martinez-Jaramillo, Computational finance, in *IEEE Computational Intelligence Society Newsletter*, pp. 8–13, August 2004
7. F. Coulmas, *The Blackwell Encyclopedia of Writing Systems* (Blackwell, Oxford, 1999)
8. J.Z. Sukkarieh, M. von Davier, K. Yamamoto, *From Biology to Education: Scoring and Clustering Multilingual Text Sequences and Other Sequential Tasks.* Educational Testing Service, Princeton, NJ, ETS Research Report No. RR–12–25, December 2012
9. R.E.I. Pardede, L.L. Tóth, G.A. Jeney, F. Kovács, G. Hosszú, Four-layer grapheme model for computational paleography. J. Inf. Technol. Res. (JITR) **9**(4), 64–82 (2016)
10. P.H. Lyu, E.W.T. Ngai, P.Y. Wu, Scientific data-driven evaluation on academic articles of low-carbon economy. Energy Policy **125**, 358–367 (2019)
11. G. Hosszú, F. Kovács, Topological Analysis of Ancient Glyphs, in Proceedings of the 2016 IEEE Trans. Syst. Man Cybern. B Cybern (SMC 2016), October 9–12, 2016, Budapest, Hungary, pp. 2248–2253
12. C. Brewster, H. Alani, S. Dasmahapatra, Y. Wilks, Data driven ontology evaluation, in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04,* Lisbon, Portugal, May 2004 (European Language Resources Association (ELRA), 2004)
13. J.A. Belmonte, M.A. Perera Betancort, C.G. García, Análisis estadístico y de grupos de las escrituras líbico – beréberes de Canarias y el Norte de África, in *VII Congreso de Patrimonio Histórico. Inscripciones Rupestres y poblamiento del Archipiélago canario*, 6–8 October 2010, Arrecife de Lanzarote
14. C.M. Skelton, Borrowing, character weighting, and preliminary cluster analysis in a phylogenetic analysis of the ancient Greek dialects. Indo-Eur. Linguist. **3**(1), 84–117 (2015)
15. P.Z. Revesz, Bioinformatics evolutionary tree algorithms reveal the history of the Cretan script family. Int. J. Appl. Math. Inf. **10**, 67–76 (2016)
16. P.Z. Revesz, M.P. Rashid, Y. Tuyishime, The design and implementation of AIDA: ancient inscription database and analytics system, in*Proceedings of the 23rd International Database Applications & Engineering Symposium, IDEAS'19,* Athens, 2019, pp. 35:1–35:6
17. M. Lidzbarski, Der Ursprung der nord-und südsemitishcen Schrift, in *Ephemeris für semitische Epigraphik. Vol. I. 1900–1902*, ed. by M. Lidzbarski (Ricker, Giessen, 1902), pp. 109–136
18. M. Bernal, *Cadmean Letters. The Transmission of the Alphabet to the Aegean and Further West before 1400 B.C.* (Eisenbrauns, Winona Lake, IN, 1990)

19. G. Hosszú, The Rovas: a special script of the Central and Eastern European Languages. Acta Philologica **44**, 91–102 (2013)
20. D.D. Vasil'ev, The Eurasian areal aspect of Old Turkic written culture. Acta Orientalia Academiae Scientiarum Hungaricae **58**(4), 323–330 (2005)
21. L. Marsadolov, The Cimmerian Traditions of the Gordion Tumuli (Phrygia). Found in the Altai Barrows (Bashadar, Pazyryk), in *Kurgans, Ritual Sites, and Settlements. Eurasian Bronze and Iron Age*, ed. by J. Davis-Kimball, E.M. Murphy, L. Koryakova, L.T. Yablonsky (Archaeopress, Oxford, 2000), pp. 247–258.
22. P.D. Ashlock, The uses of cladistics. Annu. Rev. Ecol. Syst. **5**, 81–99 (1974)
23. G. Babayarov, The imperial titles on the coins of the Western Turkic Qaghanate, in *History of Central Asia in Modern Medieval Studies (In Memoriam of Professor Roziya Mukminova)* (Yangi Nashr, Tashkent, 2013), pp. 330–348

A Relevance Study of Economic Time Series Data

Lei Han, Wei Cui, and Wei Zhang

Abstract Time series research has always been an important field in economics, finance and management. Econometrics and time series data mining are important means of time series research. This paper starting from the research on the correlation of macroeconomic data, studies the lag of macroeconomic variables and process of long-term equilibrium and short-term volatility from the perspectives of time series data mining and econometrics. Taking CPI index (CPI), money supply (M2) and total investment in fixed asset (INV) as examples to exam the above hypothesis, the result shows that there is a lag effect between CPI, M2 and INV, and there is a long-term equilibrium relationship.

Keywords CPI index · Money supply · Total investment in fixed asset · Lag · Long-term equilibrium · Short-term volatility

1 Introduction

Extract favorable information from time series data is not only profitable for enterprises, but also helps prepare preventive measures in advance for possible risks.

In the field of data mining similarity measurement, there are two common ways to measure the distance: Euclidean distance and dynamic time warping (DTW) distance. Chen pointed out that [1] dynamic time warping is helpful to solve the problem of similarity measurement when time series data with unequal length, support the shift of time series, and a flexible way deal with multi-phase time series data. Li et al.'s research shows that [2] dynamic time warping improves the calculation efficiency and

L. Han · W. Cui (✉)
School of Economics and Management, China University of Geosciences, Beijing, China
e-mail: fxcuiwei@sina.com

L. Han
e-mail: 2362347842@qq.com

W. Zhang
Digital China Group Co. Ltd, Beijing, China
e-mail: 3469988369@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_29

the accuracy of similarity matching measures., however in the process of measurement, the matching process needs to be optimized, which does not meet the triangular inequality. When DTW is used in the reduced dimension situation, it should meet the lower bound requirements to reduce the occurrence of missing reports. Yan pointed out that [3] improving the lower bound function can effectively reduce the search time. Keogh's LB_Keogh method is a measurement method to meet the lower bound requirements [4]. Ratanamahatana et al. [5] pointed out that LB_Keogh method is one of the most rapid methods for similarity measurement of time series data, which can also effectively improve the accuracy of time series data classification.

In the field of econometrics, the commonly used econometric model is the vector auto regressive (VAR) model, Shen et al. [6] found VAR model has good econometric feature, so it has been put forward and widely used, but there are also several cons, mainly focus the impact of lag period on the estimated parameters and data length, not strictly following the economic theory, parameter instability, limited number of variables to be processed, etc.. For the above cons, some scholars has improved VAR model.

The most important macroeconomic data including CPI(reflect the changes in the price levels of consumer goods and service items generally purchased by residents), M2(the sum of cash and deposits in circulation at a certain point), INV(the workload of construction and purchase of fixed assets in monetary terms), etc. Peng and other [7] empirical studies shows that the positive regulatory effect of money supply on various types of inflation is mainly concentrated on the short term, while the long-term positive regulatory effect is weak. Han [8] used cointegration analysis and error correction model to explore the relationship between M2 and CPI, and found that both short-term and long-term, M2 and CPI showed a positive relationship. Li [9] believes that M2 has obvious lag to CPI through the autoregressive distribution lag model. Qin [10] analyzed the relationship between money supply and inflation through VAR, ADF unit root test, Granger causality test, and VAR model, and concluded that money supply in China is not the cause of inflation. Lin [11] used the Archimedes Copula function and Ellipse Copula function to analyze the correlation and tail characteristics between M2 and CPI and found that the trend changes in the growth rates of M2 and CPI are consistent.

This paper selects three macroeconomic variables: CPI, M2 and INV to study the correlation among them. By analyzing the short-term change relationship among them, we can know the possible future economic condition in advance.

Table 1 Study variables

Research variable	Pronoun	Select time	Source
Consumer price index	CPI	January 2001 to December 2016	National bureau of statistics
Money supply	M2		
Total investment in fixed asset	INV		

2 Research on the Correlation of Macroeconomic Data of Time Series Shape

2.1 Data Selection

Take CPI, money supply and total investment in fixed asset as the specific research object, and study the correlation among the three variables. Table 1 shows the specific information of the research variables.

In order to avoid the influence of different variable units on the research results, the CPI, M2 and the INV are selected the monthly based data for research.

2.2 Study on the Correlation Between CPI, M2 and INV

Table 2 shows the correlation coefficients of the three research variables, and Fig. 1 shows the trend of the three research variables. By observing the correlation coefficient between CPI M2 and INV, we found that the correlation between the three variables was not consistent with the general acknowledgement. Generally the increase of M2 and INV will lead to the increase of CPI. After consideration, this paper believes that there is a lag in the correlation between the three variables, that is, the effect of money supply and investment on CPI is not reflected in the current period, but after they change for a period of time, CPI will show obvious changes. Therefore, we compared the correlation coefficients between CPI, M2 and INV in different lag periods to study the correlation among them.

As can be seen from Table 3 and Fig. 2, with the increasing of lag period, the correlation between CPI and M2 increases first and then decreases, showing a analogous parabola condition, and when lag periods number $n = 19$, the para-curve reaches

Table 2 Correlation coefficients of CPI, M2 and INV

	CPI_RATE	M2_RATE	INV_RATE
CPI_RATE	1.000000		
M2_RATE	−0.169848	1.000000	
INV_RATE	0.141891	0.510823	1.000000

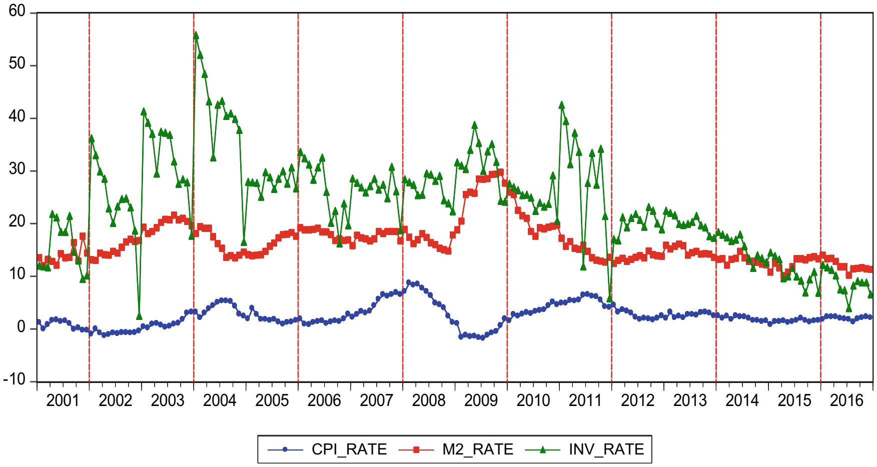


Fig. 1 Linear trend graph of CPI, M2 and INV

Table 3 CPI, M2, and INV correlations change with lag period

Lag period n	R (cpi_m2)	R (cpi_inv)	Lag period n	R (cpi_m2)	R (cpi_inv)
0	−0.1698	0.1419	16	0.4695	0.1482
1	−0.1041	0.1430	17	0.4666	0.1262
2	−0.0459	0.1576	18	0.4673	0.1164
3	0.0278	0.1713	19	0.4677	0.1202
4	0.0892	0.1911	20	0.4604	0.1193
5	0.1594	0.2155	21	0.4436	0.1379
6	0.2167	0.2272	22	0.4193	0.1424
7	0.2702	0.2195	23	0.3947	0.1492
8	0.3197	0.2143	24	0.3610	0.1464
9	0.3581	0.1853	25	0.3309	0.1566
10	0.3910	0.1720	26	0.2838	0.1479
11	0.4160	0.1717	27	0.2359	0.1602
12	0.4327	0.1707	28	0.1834	0.1787
13	0.4391	0.1759	29	0.1268	0.1755
14	0.4561	0.1767	30	0.0685	0.1499
15	0.4580	0.1668			

the peak. In general, the correlation between CPI and INV increases first and then decreases with the increase of lag. When n = 7, the correlation reaches the maximum value of positive correlation. As shown in Fig. 2, when n = 7, the correlation coefficient of CPI and M2 roughly intersects correlation coefficient of CPI and INV, and

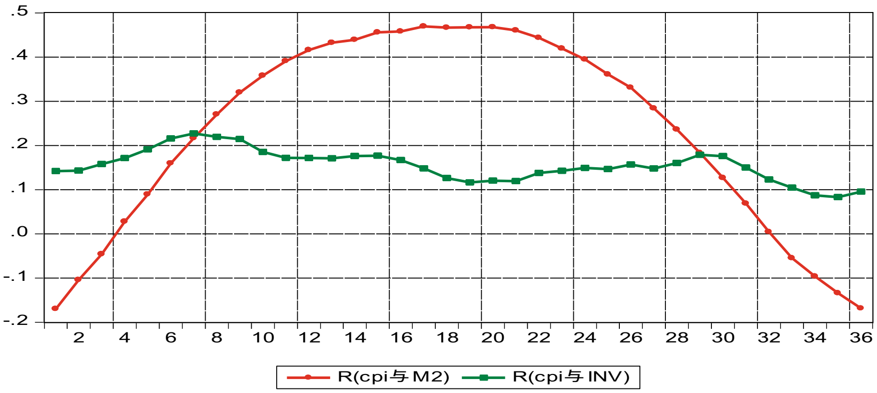


Fig. 2 Correlation of CPI, M2 and INV with lag period

the correlation coefficients are (0.2702, 0.2195), respectively. When n = 29, the two correlation coefficients intersect again, and the values are (0.1268, 0.1755).

According to the above, the lag period when the correlation coefficient between CPI and M2, CPI and INV reaches the maximum value is 19 and 7 periods respectively, i.e. 19 months and 7 months. By comparing the correlation coefficient values in these two lag periods in Table 3, it can be seen that CPI will start to show its response to the changes of M2 and INV in about 7 months.

3 Correlation Study of Time Series Macroeconomic Data

Through the research on the correlation among CPI, M2 and the INV, it is founded that there is a lag correlation among the three variables. Then we studied the correlation of macroeconomic data from two aspects of long-term equilibrium and short-term volatility. Taking CPI, M2 and INV as the specific research object, using Eviews 8.0 to test the time series shaped macroeconomic data with ADF unit root test and Johansen co integration test. ADF unit root test is mainly used to judge whether the time series is stable, Johansen's co-integration test can determing that there is a long-term equilibrium relationship between time series.

3.1 Research on Long-Term Equilibrium of CPI, M2 and INV

a. *Unit Root Test of CPI, M2 and INV*

Using ADF unit root test on CPI, M2 and INV, taking AIC criterion to conduct ADF unit root test on the original time series. It is found that the original time series is an unstable time series. The results of ADF unit root test after the first-order difference are shown in Table 4. At this time, the *P* value of all variables can reject the original hypothesis at the level of 1 and 5%, that is to say, it reaches a stable state.

Figure 3 shows the trend of each time series after the first-order difference. It can be seen from the figure that the time series of each variable has become more stable after the first-order difference, and all of them showing significant fluctuations in 2004, 2008 and 2011.

b. *Cointegration Test of CPI, M2 and INV*

According to the above unit root test results, CPI, M2 and INV are all sequences. Since they are all first-order single integer time series, Johansen co-integration test

Table 4 Results of ADF test after first-order difference

Variable	1%	5%	t-statistic	*P* value
D_CPI_RATE	−3.4674	−2.8777	−4.6623	0.0002
D_M2_RATE	−3.4674	−2.8777	−3.9991	0.0018
D_INV_RATE	−3.4676	−2.8778	−5.1675	0.0000

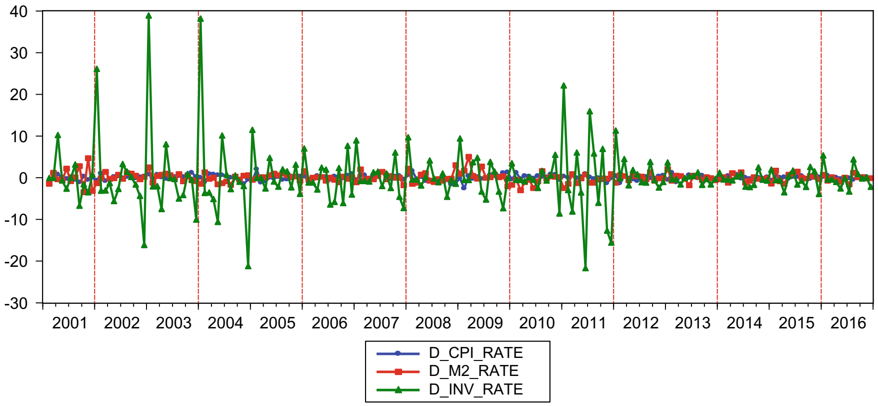


Fig. 3 First-order differential post-trend graph of CPI, M2, and INV

Table 5 Characteristic root checklist for CPI, M2, and INV

The null hypothesis: the number of cointegration vectors	Characteristic root	Statistics η	*P* value
0 Cointegration vectors	0.126540	30.09983	0.0461*
At least one cointegration vector	0.030194	6.017666	0.6935
At least two cointegration vectors	0.003143	0.560261	0.4542

* indicates that the P value is less than 0.05

Table 6 Maximum eigenvalue checklist for CPI, M2, and INV

The null hypothesis: the number of cointegration vectors	Characteristic root	Maximum eigenvalue ξ	*P* value
0 Cointegration vectors	0.126540	24.08216	0.0186*
At least one cointegration vector	0.030194	5.457404	0.6833
At least two cointegration vectors	0.003143	0.560261	0.4542

* indicates that the P value is less than 0.05

Table 7 Cointegration vector tables for standardized CPI, M2, and INV

CPI_RATE	M2_RATE	INV_RATE
1.000000	−0.301107	0.053092

can be performed to verify whether there is a long-term equilibrium relationship between the research variables.

The results of Johansen co-integration test on CPI, M2 and INV are shown in Tables 5, 6 and 7. Table 5 is the test results of characteristic root trace test statistics, and Table 6 is the test results of maximum eigenvalue. It can be seen that both test methods reject the original hypothesis, indicating that there is a cointegration vector between CPI, M2 and INV. Table 7 shows the co-integration coefficients of CPI, M2 and INV after standardization. Therefore, the co-integration coefficient matrix *B* is:

$$B = (1.000000 - 0.3011070.053092)$$

According to the analysis above, there is a cointegration vector for CPI, M2, and INV, which has a long-term equilibrium relationship.

3.2 Short-Term Fluctuation of CPI, M2 and INV

It can be seen from the above that there is a long-term equilibrium relationship between CPI, M2 and the INV. In the next section, we need to establish a short-term volatility model of time series economic data, taking CPI, M2 and the INV as the specific research object, and object, and analyzes its analyzes its short-term volatility process.

a. *Short Term Volatility Model of Macroeconomic Data Based on Time Series*

The short-term volatility model combines the dynamic time warping algorithm, sets the distance threshold, divides the time series into time series segments, uses the least-square method to establish the regression model for each time series segment, and studies the short-term volatility process through the parameter change of the time series segment regression model. The specific steps are as follows:

- Set the distance threshold Δ, use the dynamic time warping algorithm to match the similarity of time series, and then determine the segment length of the sequence and partition.
- The regression model is established for each time series segment, and the least square method is used for parameter estimation. The regression model for each segment is as follows:

$$Y_t = \alpha_t Q_t + \beta_t S_t + b_t + \varepsilon_t \tag{1}$$

  In the formula, α_t, β_t and b_t are the regression parameters under the t-th time series segment, α_t and β_t respectively represent the degree of contribution of Q_t and S_t to Y_t in the t-th segment, and ε_t represents the residuals under the t-th time series segment.
- Record and form a new time series A and B respectively, expressed as,, make corresponding statistical description for the two time series, present them in graphic form, further analyze their characteristics, find special points, and analyze their significance to dependent variables.

b. *Analysis of the short-term volatility process of CPI, M2 and INV*

- *Characteristic Analysis of Regression Parameters*

Through the establishment of the short-term volatility model above, 27 sets of regression parameters are obtained, as shown in Table 8. As can be seen from Table 8, the residual and correlation coefficients of the regression model under 27 time series segments are reasonable, in correlation coefficients most of numbers are at the level of 0.6. Therefore, as the threshold of segment size, it is reasonable, and the regression model under each segment is also reasonable. From Fig. 4, it can be seen that the regression parameter values of money supply and investment in fixed assets are not synchronized.

Table 8 Parameter values under each time series segment

Number of groups	α_t	β_t	b_t	ε_t	R^2
1	0.1768	0.0819	**−2.5659**	0.9988	0.54
2	**−0.1035**	**−0.0302**	1.8068	0.6258	0.44
3	0.1987	**−0.0446**	**−2.8022**	0.0310	0.89
4	0.5857	**−0.0043**	**−10.3273**	0.5910	0.74
5	**−0.0238**	**−0.0647**	3.4025	0.0857	0.74
6	**−0.9085**	**−0.0386**	21.7936	1.3874	0.63
7	**−0.2518**	**−0.0325**	9.9165	1.8865	0.61
8	**−0.7438**	0.0037	12.9625	2.6549	0.35
9	**−0.1826**	**−0.0843**	6.9409	0.1844	0.63
10	0.4987	0.0896	**−10.9366**	0.6009	0.27
11	**−0.6056**	**−0.0026**	12.1925	1.1690	0.42
12	1.0140	0.1117	**−16.9554**	3.9447	0.50
13	**−0.8485**	0.1645	17.9769	3.1096	0.60
14	0.6662	0.5668	**−21.1130**	13.1727	0.63
15	**−0.2369**	0.0909	1.6679	2.5309	0.44
16	0.0425	0.0471	**−4.3243**	0.1762	0.23
17	**−0.6323**	**−0.1585**	23.2194	0.0159	1.00
18	0.3106	**−1.3011**	29.4676	0.3191	0.79
19	0.7801	0.0575	**−12.2616**	1.6122	0.57
20	**−0.1505**	**−0.0517**	9.5718	0.7411	0.76
21	**−1.5170**	**−0.0956**	24.9975	2.0651	0.75
22	0.1427	**−0.1131**	2.4954	1.3148	0.15
23	**−0.3501**	0.0435	6.9412	0.6022	0.33
24	**−0.0295**	0.0982	0.9598	0.4324	0.47
25	0.2668	**−0.1128**	**−0.3414**	0.1644	0.63
26	**−0.8541**	0.1478	11.7767	0.3638	0.47
27	0.2012	**−0.0102**	**−0.2757**	0.6188	0.26

The significance of bold indicates that the value is less than 0

- *Process from Short-term Volatility to Long-term Equilibrium Adjustment Analysis*

The above verifies the existence of long-term equilibrium in the three macroeconomic time series. The short-term volatility model has been constantly changing and adjusting towards long-term equilibrium. The description of the change degree of regression parameters between each segments are shown in Table 9 and Fig. 5 respectively.

According to Table 9 and Fig. 5, the change process of regression parameters of M2 and INV changes rapidly at several special time points, while at other time

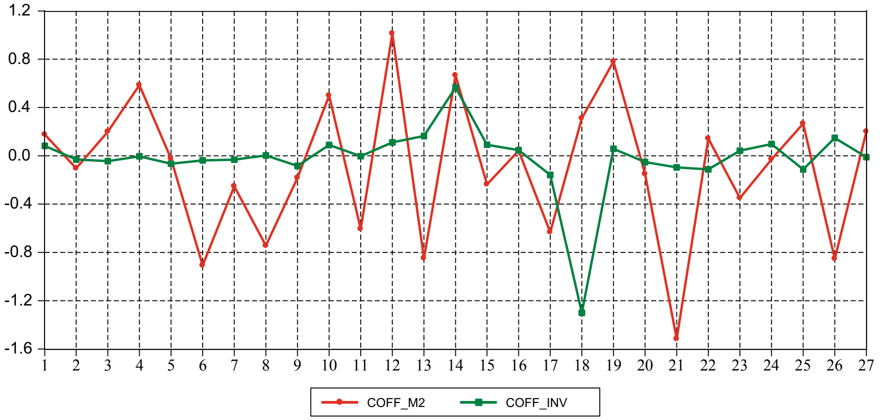


Fig. 4 Trends in the regression parameters of M2 and INV

Table 9 Rate of change of M2 and INV regression parameters

No of groups	Rate of change of α_t	Rate of change of β_t	No of groups	Rate of change of α_t	Rate of change of β_t
2–1	−2.2907	−2.1353	15–14	−0.5153	−2.3073
3–2	−0.8483	−0.7121	16–15	−8.9562	2.5330
4–3	3.9598	−2.3587	17–16	1.8853	−1.1765
5–4	−1.1718	−1.1146	18–17	−0.9905	−0.0671
6–5	−1.4019	1.9282	19–18	−5.2583	−0.4785
7–6	−15.4612	10.8222	20–19	−2.3560	−1.5112
8–7	−2.1515	−1.1971	21–20	3.2678	−26.2437
9–8	−1.1223	−2.4541	22–21	−1.0204	−0.9802
10–9	−27.8044	−3.3144	23–22	13.1606	−19.0151
11–10	−0.4036	−1.9991	24–23	−4.1817	−2.0024
12–11	−0.9046	−1.3466	25–24	−2.0778	−0.2806
13–12	26.7035	−2.3460	26–25	−1.0103	−1.6314
14–13	−0.8300	−2.4582	27–26	−7.6415	−1.0790

points, the changes tend to be stable. These special time points are shown in Table 11 respectively.

Combined of Table 10, Fig. 5 and Table 11, it can be found that when theM2 and the regression parameters of the INV change rapidly, the trend of CPI is showing as decreasing first and then increasing.

From Tables 9 and 10 and Fig. 5, it can be seen that after the CPI reached a highest point in 2008, i.e. after the 15th and 16th groups, the change rate of M2 and INV showed a significant increase trending. After that, the CPI decreased correspondingly,

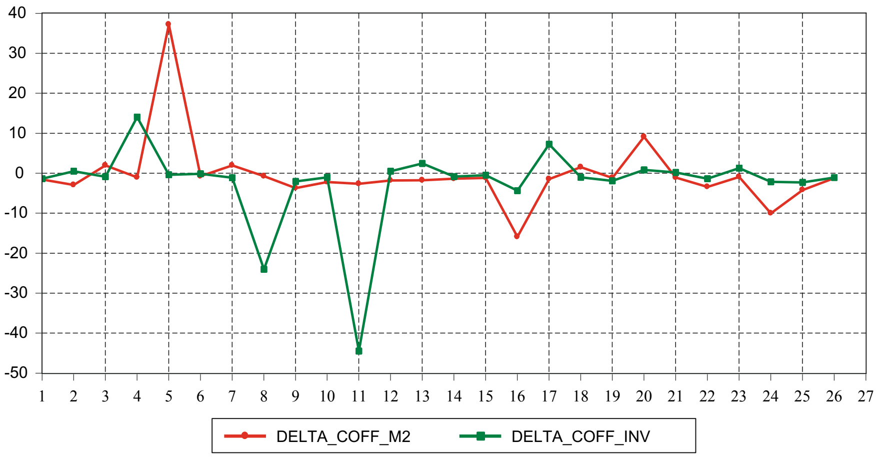


Fig. 5 M2 and INV regression parameter change rate trend graph

Table 10 Time point of change rate of M2 and INV regression parameters

No of groups	α_t Special point of rapid rate of change	Corresponding time point	β_t Special point of rapid rate of change	Corresponding time point
1	Group 5–Group 6	2003.04–2004.02	Group 4–Group 5	2002.10–2003.08
2	Group 7–Group 8	2004.03–2005.06	Group 8–Group 9	2004.11–2005.12
3	Group 16–Group 17	2009.05–2009.12	Group 11–Group 12	2006.07–2007.07
4	Group 20–Group 21	2011.01–2012.08	Group 13–Group14	2007.08–2008.12
5	Group 22–Group 23	2012.09–2013.12	Group 16–Group17	2009.05–2009.12
6	Group 24–Group25	2014.01–2015.06	Group 17–Group18	2009.09–2010.05

Table 11 Rate of change of CPI

Time	CPI change
2003.04–2003.08	1.0–0.9↓
2004.11–2005.06	2.8–1.6↓
2009.05–2009.12	−1.4–1.9↑

maintaining at about 3%, indicating that the control of the supply and investment of M2 and INV can be effective To restrain inflation.

- *Characteristic Analysis of Key Regression Points*

As shown in Fig. 4, in the analysis of the short-term fluctuations of the three macroeconomic time series of CPI, M2 and INV, it is obvious that each time series segment group shows sensitive to regression parameters of M2 and INV are group 14 and group 16 respectively, i.e. from May to December in 2008 and from May to August in 2009.By observing the value changes of CPI in these two time series segments and the segments before and after them, it can be found that CPI has been increasing from 3.3% to 8.5% at April 2008, which means that inflation will occur if M2 and INV can no longer be effectively controlled. According to the fact global economic crisis broke out in 2008, shows from January 2007 to July 2007, the 12th group of M2 in this time segment and the regression parameters of the INV predict the possible inflation.

With the above analysis, we can draw a conclusion that the influence of M2and INV on the CPI has a lag effect, which further indicates that it can be achieved by the change of M2 and INV can predict the possible change trend of CPI in advance.

4 Conclusion

This paper studies the correlation of time series shaped macroeconomic data from the aspects of lag correlation, long-term equilibrium and short-term fluctuation, and takes CPI,M2 and INV as the specific research object, and discusses the relationship between the three variables.

In the research of lag correlation, by comparing the changes of correlation coefficient in different lag periods, we can find that there is a lag effect between CPI, M2 and INV, with a lag period of about 7 months.

In the aspect of long-term equilibrium research, through ADF unit root test and Johansen co-integration test, it is concluded that CPI, M2 and INV are all first-order single integration sequences, and there is a co-integration vector, which has a long-term equilibrium relationship.

In the study of short-term volatility, combining dynamic time warping and the least square method, the paper constructs a short-term volatility model, and concludes that the future trend of CPI can be obtained through the analysis of key regression segments, which is helpful to predictive of the occurrence of inflation.

References

1. H.Y. Chen, C.H. Liu, B. Sun, A survey of similarity measures in time series data mining. Control Decis. **32**(1), 1–11 (2017)
2. Z.X. Li, F.M. Zhang, K.W. Li, Multi-time sequence pattern matching method based on DTW. Pattern Recogn. Artif. Intell. **24**(3), 425–430 (2011)

3. Z. Yan, W.J. Su, C.C. Yu, Z.J. Wu, An improved DTW similarity search method. Comput. Simul. **36**(2), 232–236+270 (2019)
4. E. Keogh, C.A. Ratanamahatana, Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**(3), 358–386 (2005)
5. C.A. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints, pp. 11–22 (2004)
6. Y. Shen, S.Z. Li, X.T. Ma, The evolution and latest development of VAR macro-econometric model—based on the expansion of SMIs research results of nobel prize winner in 2011. Quant. Econ. Technol. Res. **10**, 150–160 (2012)
7. H.F. Peng, H.Y. Zhao, Correlation between money supply and inflation: based on time and frequency dual perspective. Syst. Eng. Theory Pract. **36**(8), 1905–1917 (2016)
8. S.Q. Han, The influence of broad money supply on the consumer price index. Mark. Res. **4**, 43–44 (2017)
9. M. Li, Empirical analysis of China's money supply (M2) and consumer price index (CPI)-based on autoregressive distribution lag model. Times Finance **8**, 8–10 (2017)
10. Q.X. Qin, Analysis of the relationship between money supply and inflation. Technol. Ind. Across Straits **2**, 44–45 (2019)
11. W. Lin, J. Zhou, Research on the relationship between broad money supply and CPI changes—an empirical analysis based on copula function. Price Theory Pract. **11**, 70–73+100 (2019)

Deciphering Historical Inscriptions Using Machine Learning Methods

Loránd Lehel Tóth, Gábor Hosszú, and Ferenc Kovács

Abstract This paper presents the results of a decipher historical inscriptions approach to demonstrate the application of different similarity metrics, classification and algorithm acceleration methods. Deciphering historical inscriptions is difficult in the most cases because the survived inscriptions typically contain calligraphic glyphs, grapheme errors or incomplete words. The basis of the presented methods are the geometric-topological features, which form feature vectors for each glyphs and undeciphered symbols that describes the shape of them with the numerical data. The elaborated method calculates the similarity distances of the inscriptions by the matching accuracies of the recognized graphemes through their topological feature vectors and determines the meaning of the inscription using an external dictionary of historical words. The actual version of the deciphering software is restricted for the one-word-long inscriptions. The article presents experimental results, which were processed on a real inscription. It demonstrates the efficiency of the methods. The deciphering software could be used for a paleographical research, especially in deciphering ancient hard to read inscriptions.

Keywords Computational palaeography · Deciphering ancient inscriptions · Pattern recognition · Topological features · Scriptinformatics · Similarity metrics

L. L. Tóth (✉) · G. Hosszú · F. Kovács
Budapest University of Technology and Economics, Budapest, Hungary
e-mail: lorand.toth@tungsram.com

G. Hosszú
e-mail: hosszu@eet.bme.hu

F. Kovács
e-mail: kovacsf@itk.ppke.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_30

1 Introduction

After exploring ancient inscriptions, several times the researchers found it difficult to decipher them, determine their age and style and identify their writers. The reason for this, apart from the deterioration of the writing material (wood, stone, brick, paper, etc.), is that the shapes of the letters used in writing have changed over the time. Our research is focusing on deciphering of hard to read inscriptions, using statistical methods and based on the geometric-topological features of the glyphs and undeciphered symbols.

Computational approaches to historical linguistics were proposed for half a century. Since then numerous research have been published in the topic of pattern recognition, OCR, writing recognition, deciphering ancient inscription. The variety of books, articles, journals and conferences in the international literature confirm the usefulness of this topic. Within the last decade, the new wave of computational historical linguistics has been introduced. These are automatic assessment of genetic relatedness, automatic cognate detection, phylogenetic inference and ancestral state reconstruction exploring the bigger picture of a language family's phylogeny, which are collected and discussed in [1, 2]. Our approach differs from the well-known optical character recognition (OCR) functions [3]. While in case of OCR the image processing is a common integral part of the method, we don't use image processing since we register the form shapes of the examined glyphs and symbols manually. Also, OCR's mainly focusing to recover characters from an image, while we are focusing on deciphering of the meaning of an unknown inscription, however our methods are using some common tasks with OCR like glyph segmentation, feature selection. The OCR methods are not only focused on recognizing of an ancient inscriptions, but these are used effectively for text localization and text recognition in video and picture processing applications [4].

The machine learning techniques are widely used in character recognition field if the large amount of data is available to compose training sets, this condition is not always fulfilled. Good results were achieved for handwritten Bengali numerals recognition using Convolutional Neural Networks, which is the major language spoken in the Indian subcontinent, and even the first and official language of Bangladesh [5]. In article [6] a comparison of several approaches for visual recognizing ancient inscriptions is provided. Their experiments, performed on 17,155 photos related to 14,560 inscriptions, showed that combining Fisher Vector and Convolutional Neural Network features into a single image representation resulted a very high effectiveness of correctly recognizing the query inscription, that is more than 90% of the cases.

In [7] authors proposed a method for the automatic decipherment of the lost languages. The key strength of their model lies in its ability to incorporate a range of linguistic intuitions in a statistical framework. Applied to the ancient Semitic language Ugaritic, the model deduces the correct Hebrew cognate for 60% of the Ugaritic words which has cognates in Hebrew and correctly maps 29 of 30 letters to their Hebrew counterparts. They used character level comparisons and statistical methods to reach their goals.

Another major area of the pattern recognition and data mining techniques in historical inscriptions is to discover relationships between ancient scripts (writing systems) including their possible common origin from a single root script. Paper [8] presents data mining techniques using convolutional neural networks and support vector machines to find the degree of visual similarity between pairs of symbols in eight different ancient scripts. Authors in [9] had also achieved very good results in discovering relationships of the ancient scripts. They presented a machine learning approach to explore the phenetic relations of historical scripts and their glyphs considering their topological properties and transformations in the development of the glyphs. They used different cluster analysis methods based on the similarity groups of the glyphs of the historical scripts in order to explore the phenetic relationships between these scripts. A general methodology for identifying the writer of an ancient inscription has been presented in [10], which includes methods of pattern recognition, image processing and mathematics. This methodology uses the realizations of an alphabet symbol appearing in an inscription and compares them based on the original criteria. Identifying the writer of an ancient inscription is important for Archaeometry and History, since it helps to recognize the origin of its content. A novel statistical criteria for deciding have been developed to check whether two inscriptions are made by the same writer or by different writers.

The similarity measurement is used for text classification and cluster analysis is also a known method to find the common features and to compute the similarity between two documents [11]. Our solution is to perform the similarity measurement between a known glyph and an unknown symbol using their geometric-topological feature vectors.

Another interesting area is where researchers rely on 3D modeling algorithms to digitize and decipher historical inscriptions. Authors in [12] proposed a novel framework for efficient 3D reconstruction of inscriptions and for statistical analysis of their reconstructed surfaces. They were using a shape-from-shading technique to reconstruct in 3D the shape of the inscribed surfaces, which were segmented into a smaller box-shaped regions containing single character. These characters were classified into groups of the same characters or symbols and then an atlas letter shape was created for each character. Using the atlases an automated analysis of the inscribed letters was performed. This framework could be effectively used for the study of the variations of the lettering techniques within an inscription or a set of inscriptions. In article [13] the author presented an algorithm for the automated reconstruction and visualization of damaged ancient inscriptions using a hybrid approach that combined advantages of 2D and 3D analytical techniques.

Our research is focusing to decipher hard to read or undeciphered one-word long parts of Székely-Hungarian Rovash inscriptions from fifteenth to seventeenth century [14]. The Székely-Hungarian Rovash script belongs to the Rovash (pronounced “rove-ash”, other spelling: Rovas) script family with other members like the Turkic Rovash (also called Turkic Runic), Steppean Rovash, and Carpathian Basin Rovash. The main issue was that the historical finds are available with limited numbers [15].

2 Description of the Algorithms

Its first step is the registration of the topological features of an undeciphered symbol. Different topological features of the examined glyph were introduced by the researchers described in [16–19]. In our interpretation the topological features describe a glyph or a symbol, which can be e.g. closed loop, vertical line, horizontal line, endpoint, etc., which can be visually recognized and identified.

Our method is built up from two parts. The first one is called SID-Preprocessing algorithm while the second one is called SID-Main algorithm. The SID-Preprocessing algorithm collects the possible cognate glyphs for each symbol by minimizing the similarity distances of the topological vectors. Thereby a set of a known glyphs belong to each symbol. The glyphs have not only topological features but also transliteration values and sound values, which is important in the means of a glyph.

The historical words dictionary contains the words in their sound values.

We introduced the transliteration values in article [20]. One transliteration value can belong to a different glyphs and different sound values [21]. Transliteration values were used for dimension reduction purposes in our case. One approach for phoneme-to-grapheme transliteration for speech recognition systems is presented in article [22].

As a next step the SID-Preprocessing algorithm creates the set of glyph strings as a combination of the cognate glyphs symbol by symbol. This step is important because the complete uniformity of the most cognate glyph and the examined symbol is not guaranteed. We are able to determine the quantity of the most cognate glyphs being processed. The set of known glyphs could change time by time during the test runs based on the choice of a different initial input parameter settings.

Before making combinations of the previously selected glyph set, the algorithm replaces the glyphs to their transliteration values within one set, then merge them together, eliminating duplications. As a result, the same set of glyph could contain different glyph shapes with the same sound values. Executing the combinations with these smaller sets creates transliteration strings. Finally, the algorithm replaces the transliteration values to sound values, since more than one sound value can belong to the same transliteration value. This creates strings of all possible combinations of sound values that can now be searched in the historical word dictionary.

The hits in the historical dictionary will be the output of the SID-Preprocessing algorithm and input of the SID-Main algorithm. We determine in this phase that the hits should contain the most relevant deciphering of the input symbol string, but we do not yet know the amount of similarity. The SID-Main algorithm is responsible for determining this similarity level using quantitative description and select the most similar glyph determination for the undeciphered symbol string. Thus we have the potential historical words set that will be processed one by one with the following operations.

The algorithm replaces the sound values to their transliteration values, therefore the transliteration values are replaced with their glyphs. With this method our algorithm generates all the possible glyph strings from their glyph variants. Finally, the topological parameter vectors of the generated glyph strings are compared with the parameter vectors of symbol string using two basic types, but very effective similarity metrics the Hamming - and Euclidean similarity metrics [23, 24], and as the output of the SID-Main algorithm the set of the most relevant glyph strings will be selected in ascending order of their similarity values.

3 Description of the Applied Methods

3.1 Classification Method

We introduced an accelerator method into the algorithm. The glyphs and the undeciphered symbols of ancient inscriptions have grouped using cluster analysis methods on the topological parameters [25]. This grouping method were called “T-classifier” (T stands for “topological”) according to the input of the algorithm, which are topological parameters of the symbol or glyph. Each time a new symbol or glyph occurs, the cluster analysis method has to be performed to find the proper group where the glyph/symbol belongs to.

In addition to the “T-classifier” another classification method was implemented in the software. It is a Visual classifier called “V-classifier”, that is based on heuristic method [26]. This classification was performed by the authors manually based on the visual attribution of the glyphs and symbols. The output of the “V-classifier” differs from the output of “T-classifier” in some cases. Contrary to the “T-classifier” we did not use pattern recognition and machine learning techniques during the “V-classifier” preparation. The classes were formed in an intuitive way using visual aspects of the glyphs and symbols.

In case a new symbol is registered in the database, it’s classification fields “T-” and “V-classifier” are defined. If the user sets this parameter in the setting window, the examined undeciphered symbol will be compared only to the set of glyphs which are in the same class, group. This acceleration method gives relevant speed to the algorithm during the application. The disadvantage is that if a symbol is misclassified, the SID-Preprocessing algorithm will not find the potential matches in historical word database, therefore the SID-Main algorithm will not get the input data. The potential negative result of this disadvantage will be corrected using the “Levenshtein distance” option.

3.2 Levenshtein Distance

The final step of the SID-Preprocessing is using Levenshtein similarity metric during the comparison of the generated sound value string with the historical words [27, 28]. If this value reaches a predefined similarity threshold, the actual words will be selected from the historical dictionary as the results of the SID-Preprocessing algorithm. This threshold defines the allowed number of differences between the generated string and the historical words in the dictionary. The number of the results depends from the predefined threshold, that the user can change in the GUI of the software before running, as a variable parameter. If the threshold is quite large the SID-Preprocessing algorithm returns with more words, even if they are not enough relevant; otherwise, few precise hits will be returned to the output.

The test results proved that the SID-Preprocessing algorithm gives positive hits even in those cases when the generated sound value strings are not equal with the sound values of the words found in the historical word dictionary, which eliminates the negative effect of the misidentified symbols in the previous steps.

3.3 Circle Method

We developed a geometric-topological based method, called "Circle method" presented in article [20]. The accuracy of our deciphering algorithm has been further improved with the Circle method, which is based on three concentric circles structured around graphemes using the polar coordinate system [29] that generates cross section points with the skeletons [30] of examined glyphs and symbols. The deciphering algorithm runs on the topological feature vectors, which could be supplemented with the set of three extra features as an option. The generated features are rotation-, translation-, and size-invariant. Test runs showed that using the additional feature vector provided by Circle Method gave more accurate deciphering results during deciphering hard to read symbols, but resulted in more processing time than without using the Circle Method.

3.4 Number of Matches

"Number of Matches" parameter defines the size of the set of similar glyphs belonging to the undeciphered symbol that will be processed. If this value is set to a large number, the probability to find the corresponding glyph is high, but this also is inversely proportional to the processing time, based that the combination of the examined glyph strings will growth exponentially. It is worth keeping this value low, especially when we want to decipher longer words.

3.5 Algorithm Robustness

We introduced the "robustness level", which significantly improves the processing time of the SID-Main algorithm by selecting the examined glyph strings (describes meaningful words) in a predefined sequence from the full combination set of glyph strings. Only the pre-filtered set of string will be examined. This has less effect on finding the best results, but at the same time gives an opportunity to display different results (different glyph words with their similarity distances) on the output.

4 Application of the Methods

This section presents our software called SID (inScription IDentification). We determine the meaning of the real inscription found in a historical written material called "Kájoni-prior" and was created by János Kájoni Franciscan monk in 1673. Kájoni used an alphabet which was different in the representation of the glyphs known before, however he used the same language and sound values than his contemporaries. The original inscription was written from right to left direction, that we rewrite from left to right direction as seen in Fig. 1 and compose the glyphs and upload to our symbol database supplemented with it's topological features but without their sound values and transliteration values, as we supposed that those are undeciphered.

The experiments were carried out on a standard laptop with an Intel I5-3320M @2.60 GHz processor with 8 GB of RAM, the operating system was Microsoft Windows 10 with XAMP control panel v3.2.2 which includes the Apache and MySQL servers. The SID software were written in PHP/HTML and using MySQL database.

We performed some test runs with different setting options to determine the optimal settings, thus we got more accurate results with shorter processing time. The main screen of the SID software is showed in Fig. 2.

Using the settings illustrated in Table 1, the algorithm could not find any matches in the historical words dictionary seen in Fig. 3.

In next run we set the "Levenshtein distance" value to 3 and left the other parameters invariably listed in Table 2. The algorithm found a relevant word in the historical words dictionary "örökké = forever", which is the only one and best meaning of the undeciphered inscription seen in Fig. 5. The algorithm listed the best relevant representation form with glyph variants of this word seen in Fig. 4, and calculated the lower Hamming distance and Euclidean distance for it. The processing time was 66 s.

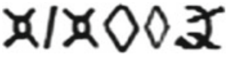
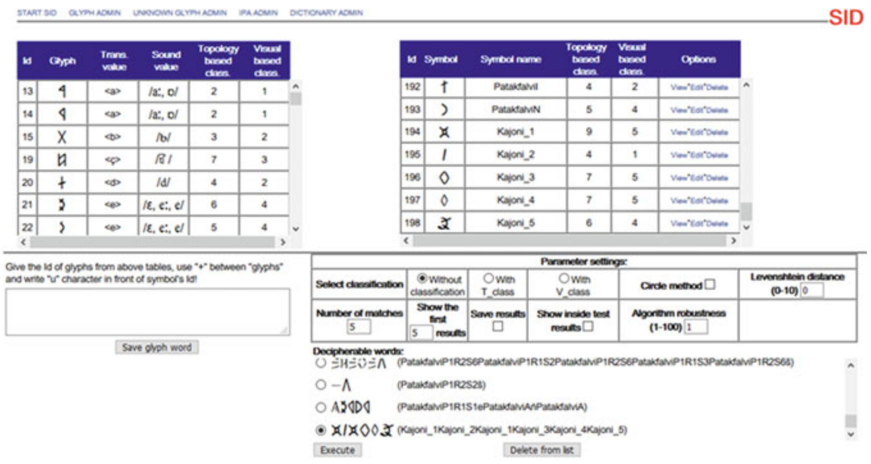


Fig. 1 Composed inscription based on the original inscription



Id	Glyph	Trans. value	Sound value	Topology based class.	Visual based class.
13	⁊	<a>	/aː, ɒ/	2	1
14	⁊	<a>	/aː, ɒ/	2	1
15	X	<b>	/b/	3	2
19	ꓘ	<ç>	/ʦ̑ /	7	3
20	∤	<d>	/d/	4	2
21	ʒ	<e>	/ɛ, eː, e/	6	4
22	)	<e>	/ɛ, eː, e/	5	4

Id	Symbol	Symbol name	Topology based class.	Visual based class.	Options
192	†	PatakfalvіI	4	2	View*Edit*Delete
193	)	PatakfalviN	5	4	View*Edit*Delete
194	Ж	Kajoni_1	9	5	View*Edit*Delete
195	/	Kajoni_2	4	1	View*Edit*Delete
196	◊	Kajoni_3	7	5	View*Edit*Delete
197	◊	Kajoni_4	7	5	View*Edit*Delete
198	ꭗ	Kajoni_5	6	4	View*Edit*Delete

Give the Id of glyphs from above tables, use "+" between "glyphs" and write "u" character in front of symbol's Id!

Save glyph word

Parameter settings:

Select classification	● Without classification	○ With T_class	○ With V_class	Circle method □	Levenshtein distance (0-10) 0
Number of matches 5	Show the first 5 results	Save results □	Show inside test results □	Algorithm robustness (1-100) 1	

Decipherable words:

○ ⫶Ħ⫶◊⫶Λ (PatakfalviP1R2S6PatakfalviP1R1S2PatakfalviP1R2S6PatakfalviP1R1S3PatakfalviP1R2S6ß)

○ –Λ (PatakfalviP1R2S2ß)

○ ΛʒᑫDᑫ (PatakfalviP1R1S1ePatakfalviAnPatakfalviA)

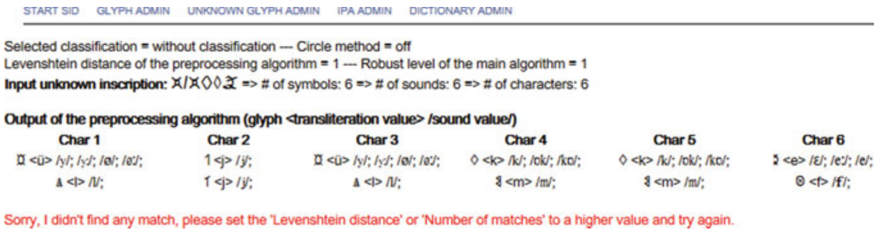
● Ж/Ж◊◊ꭗ (Kajoni_1Kajoni_2Kajoni_1Kajoni_3Kajoni_4Kajoni_5)

Execute Delete from list

Fig. 2 Main screen of the SID

Table 1 Input parameters of test-run 1

Initial input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	Without classification	Off	1	1



Selected classification = without classification --- Circle method = off
Levenshtein distance of the preprocessing algorithm = 1 --- Robust level of the main algorithm = 1
Input unknown inscription: Ж/Ж◊◊ꭗ => # of symbols: 6 => # of sounds: 6 => # of characters: 6

Output of the preprocessing algorithm (glyph <transliteration value> /sound value/)

Char 1	Char 2	Char 3	Char 4	Char 5	Char 6
Ж <ü> /y/; /yː/; /ø/; /øː/;	1 <j> /j/;	Ж <ü> /y/; /yː/; /ø/; /øː/;	◊ <k> /k/; /ɒk/; /kɒ/;	◊ <k> /k/; /ɒk/; /kɒ/;	ʒ <e> /ɛ/; /eː/; /e/;
▲ <l> /l/;	1 <j> /j/;	▲ <l> /l/;	∛ <m> /m/;	∛ <m> /m/;	⊙ <f> /f/;

Sorry, I didn't find any match, please set the 'Levenshtein distance' or 'Number of matches' to a higher value and try again.

Fig. 3 Result of the SID algorithm without any matches

Table 2 Input parameters of test-run 2

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	Without classification	Off	3	1

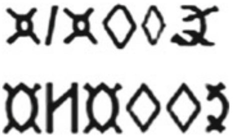
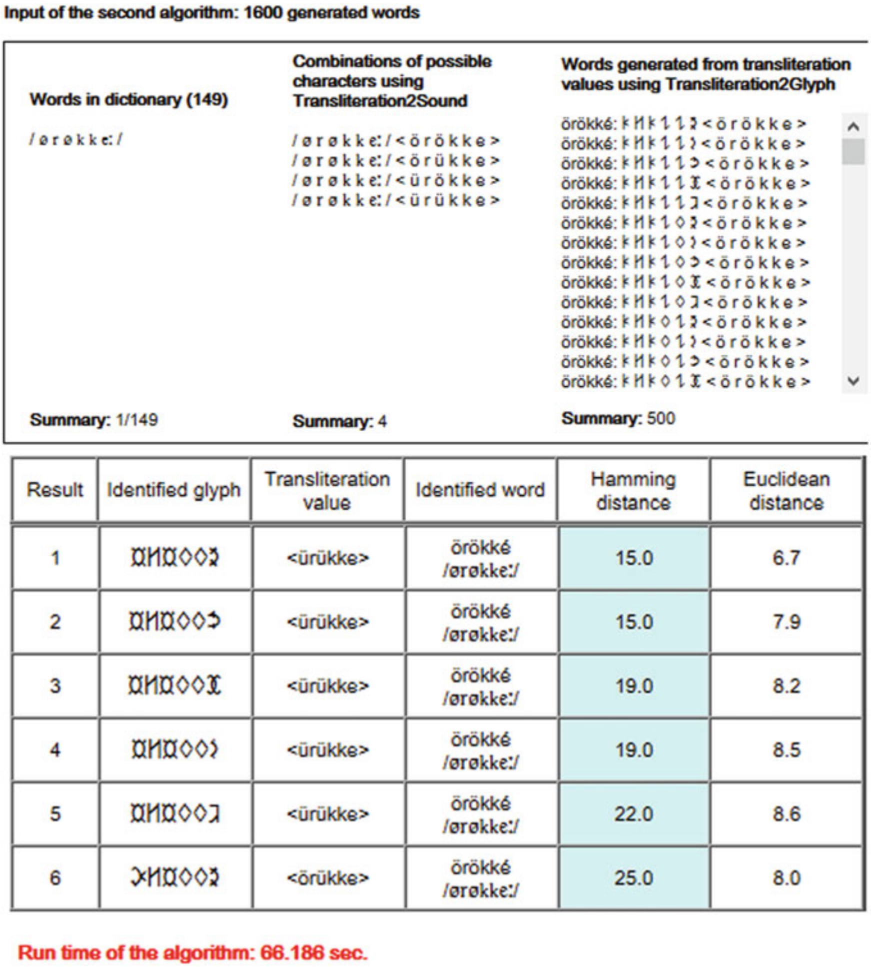


Fig. 4 Undeciphered symbol string and it's most similar deciphering



Input of the second algorithm: 1600 generated words

Words in dictionary (149)	Combinations of possible characters using Transliteration2Sound	Words generated from transliteration values using Transliteration2Glyph
/ørøkkeː/	/ørøkkeː/<örökke> /ørøkkeː/<örükke> /ørøkkeː/<ürökke> /ørøkkeː/<ürükke>	örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke> örökké: <örökke>
Summary: 1/149	Summary: 4	Summary: 500

Result	Identified glyph	Transliteration value	Identified word	Hamming distance	Euclidean distance
1		<ürükke>	örökké /ørøkkeː/	15.0	6.7
2		<ürükke>	örökké /ørøkkeː/	15.0	7.9
3		<ürükke>	örökké /ørøkkeː/	19.0	8.2
4		<ürükke>	örökké /ørøkkeː/	19.0	8.5
5		<ürükke>	örökké /ørøkkeː/	22.0	8.6
6		<örükke>	örökké /ørøkkeː/	25.0	8.0

Run time of the algorithm: 66.186 sec.

Fig. 5 Result of the SID algorithm with one match

Table 3 Input parameters of test-run 3

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	Without classification	Off	5	1

In next run we changed parameters according Table 3. The algorithm found three different relevant words in the historical words dictionary "örökké = forever", "örökre = for ever" and "öröklő = inherited", however the best meaning of the undeciphered inscription is still the "örökké = forever" word, seen in Fig. 6. As it seen the processing time of the algorithm was increased to 122 s.

In next run we changed parameters according Table 4. The algorithm still found three different relevant words in the historical words dictionary "örökké = forever", "örökre = for ever" and "öröklő = inherited" because the SID-Preprocessing algorithm does not take into consideration the effect of this method. In contrast, it seems that the deciphering list was changed. We can see new results since the SID-Main algorithm processed in larger scale the possible hits from the generated glyph strings. The best meaning of the undeciphered inscription is still the "örökké = forever" word, but with a different glyph representation mode, seen in Fig. 7. As it seen the processing time of the algorithm was decreased significantly to 7 s.

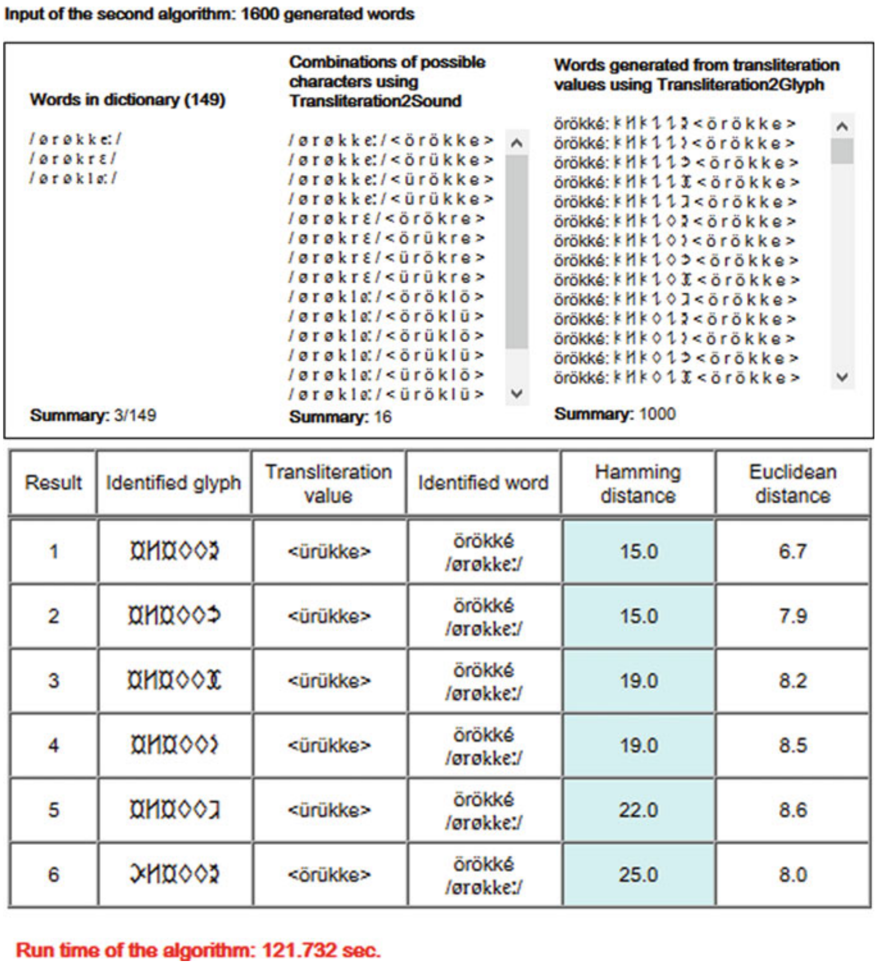
In next run we changed parameters according Table 5. The algorithm found the same three different relevant words in the historical words dictionary "örökké = forever", "örökre = for ever" and "öröklő = inherited" as found in the previous run and the final results with the same values also were the same as we received before, due to the input of the SID-Main algorithm was the same, than the previous run, seen in Fig. 8. As it can be seen, the processing time of the algorithm was increased to 23 s, as the SID-Preprocessing algorithm was performed on more cognate glyphs per symbol.

Using settings in Table 6, the algorithm could not find any matches in the historical words dictionary seen in Fig. 9.

We changed only the Classification parameter to "T-Classification" according to Table 7, however the algorithm could not find any matches in the historical words dictionary seen in Fig. 10.

In next step we changed the "Classification" parameter to "V-Classification", according to Table 8. The algorithm found a relevant word in the historical words dictionary "örökké = forever" which is the only one and best meaning of the undeciphered inscription as we also got before, seen in Fig. 11. The algorithm calculated different Hamming distance and Euclidean distance in this run compared to the Fig. 5 results, because the "Circle method" was turned "On", this step added three additional topological features to the vectors, therefore the processing time was increased to 67 s.

The performed tests proved the effects of the introduced methods that were presented in the previous sections.



Input of the second algorithm: 1600 generated words

Words in dictionary (149)

/ørøkkeː/
/ørøkrɛ/
/ørøklœː/

Summary: 3/149

Combinations of possible characters using Transliteration2Sound

/ørøkkeː/<örökke>
/ørøkkeː/<örükke>
/ørøkkeː/<ürökke>
/ørøkkeː/<ürükke>
/ørøkrɛ/<örökre>
/ørøkrɛ/<örükre>
/ørøkrɛ/<ürökre>
/ørøkrɛ/<ürükre>
/ørøklœː/<öröklö>
/ørøklœː/<öröklü>
/ørøklœː/<örüklö>
/ørøklœː/<örüklü>
/ørøklœː/<üröklö>
/ørøklœː/<ürökIü>

Summary: 16

Words generated from transliteration values using Transliteration2Glyph

örökké: ᚴИᚴ11ʒ<örökke>
örökké: ᚴИᚴ11ʅ<örökke>
örökké: ᚴИᚴ11ↄ<örökke>
örökké: ᚴИᚴ11Ӿ<örökke>
örökké: ᚴИᚴ11ℷ<örökke>
örökké: ᚴИᚴ1◇ʒ<örökke>
örökké: ᚴИᚴ1◇ʅ<örökke>
örökké: ᚴИᚴ1◇ↄ<örökke>
örökké: ᚴИᚴ1◇Ӿ<örökke>
örökké: ᚴИᚴ1◇ℷ<örökke>
örökké: ᚴИᚴ◇1ʒ<örökke>
örökké: ᚴИᚴ◇1ʅ<örökke>
örökké: ᚴИᚴ◇1ↄ<örökke>
örökké: ᚴИᚴ◇1Ӿ<örökke>

Summary: 1000

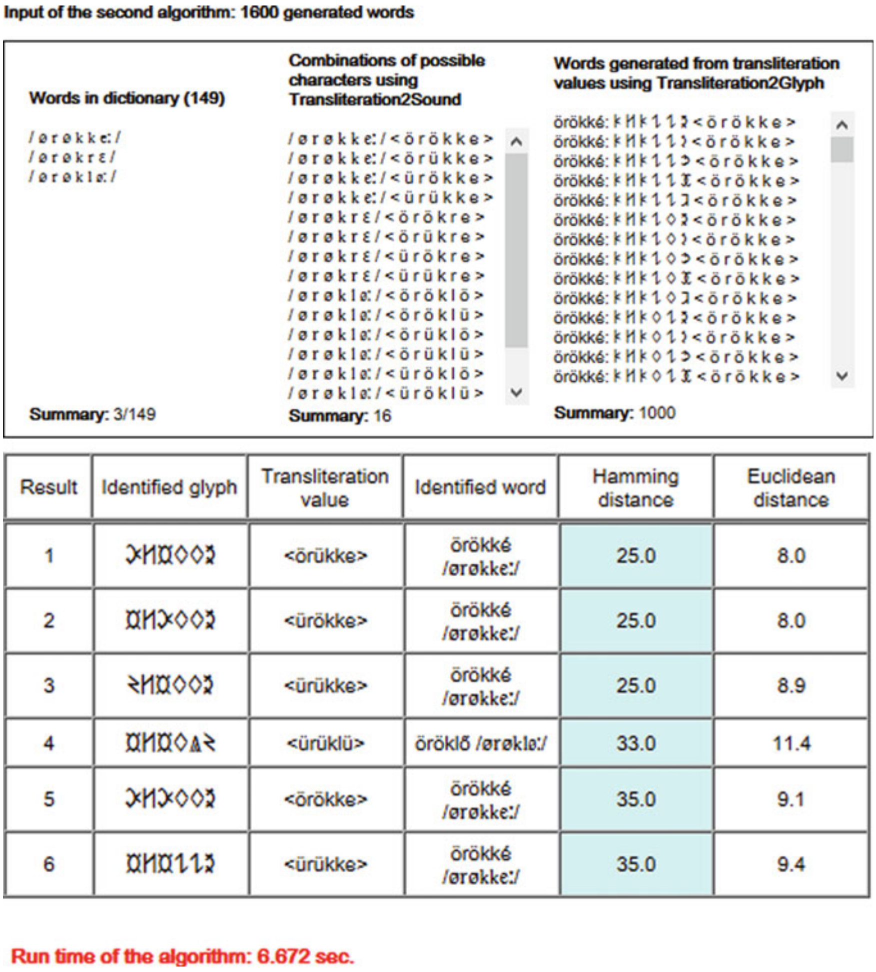
Result	Identified glyph	Transliteration value	Identified word	Hamming distance	Euclidean distance
1	ЖИЖ◇◇ʒ	<ürükke>	örökké /ørøkkeː/	15.0	6.7
2	ЖИЖ◇◇ↄ	<ürükke>	örökké /ørøkkeː/	15.0	7.9
3	ЖИЖ◇◇Ӿ	<ürükke>	örökké /ørøkkeː/	19.0	8.2
4	ЖИЖ◇◇ʅ	<ürükke>	örökké /ørøkkeː/	19.0	8.5
5	ЖИЖ◇◇ℷ	<ürükke>	örökké /ørøkkeː/	22.0	8.6
6	ϿИЖ◇◇ʒ	<örükke>	örökké /ørøkkeː/	25.0	8.0

Run time of the algorithm: 121.732 sec.

Fig. 6 Result of the SID algorithm with three matches

Table 4 Input parameters of test-run 4

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	Without classification	Off	5	25



Input of the second algorithm: 1600 generated words

Words in dictionary (149)

/ørøkkeː/
/ørøkrɛ/
/ørøkløː/

Summary: 3/149

Combinations of possible characters using Transliteration2Sound

/ørøkkeː/<örökke>
/ørøkkeː/<örükke>
/ørøkkeː/<ürökke>
/ørøkkeː/<ürükke>
/ørøkrɛ/<örökre>
/ørøkrɛ/<örükre>
/ørøkrɛ/<ürökre>
/ørøkrɛ/<ürükre>
/ørøkløː/<öröklö>
/ørøkløː/<öröklü>
/ørøkløː/<örüklö>
/ørøkløː/<örüklü>
/ørøkløː/<üröklö>
/ørøkløː/<üröklü>

Summary: 16

Words generated from transliteration values using Transliteration2Glyph

örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>

Summary: 1000

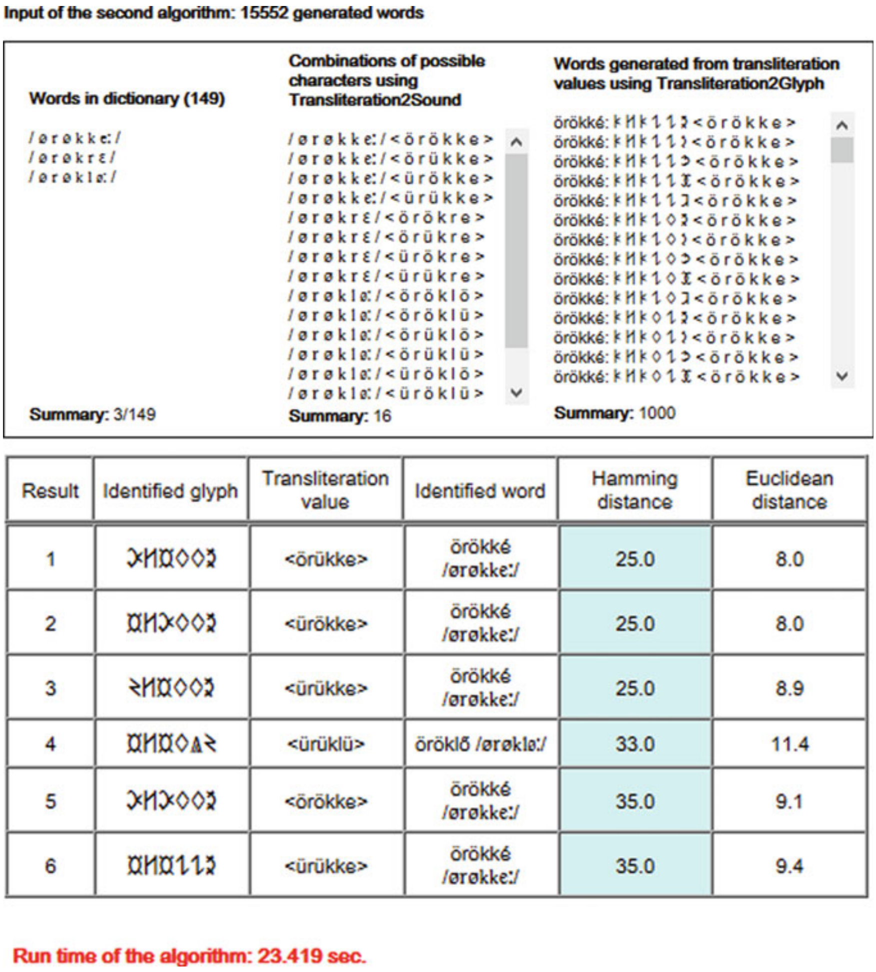
Result	Identified glyph	Transliteration value	Identified word	Hamming distance	Euclidean distance
1		<örükke>	örökké /ørøkkeː/	25.0	8.0
2		<ürökke>	örökké /ørøkkeː/	25.0	8.0
3		<ürükke>	örökké /ørøkkeː/	25.0	8.9
4		<üruklü>	öröklő /ørøkløː/	33.0	11.4
5		<örökke>	örökké /ørøkkeː/	35.0	9.1
6		<ürükke>	örökké /ørøkkeː/	35.0	9.4

Run time of the algorithm: 6.672 sec.

Fig. 7 Result of the SID algorithm with three matches and increased robust level

Table 5 Input parameters of test-run 5

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
3	Without classification	Off	3	25



Input of the second algorithm: 15552 generated words

Words in dictionary (149)

/ørøkkeː/
/ørøkrɛ/
/ørøkløː/

Summary: 3/149

Combinations of possible characters using Transliteration2Sound

/ørøkkeː/<örökke>
/ørøkkeː/<örükke>
/ørøkkeː/<ürökke>
/ørøkkeː/<ürükke>
/ørøkrɛ/<örökre>
/ørøkrɛ/<örükre>
/ørøkrɛ/<ürökre>
/ørøkrɛ/<ürükre>
/ørøkløː/<öröklõ>
/ørøkløː/<öröklü>
/ørøkløː/<örüklõ>
/ørøkløː/<örüklü>
/ørøkløː/<üröklõ>
/ørøkløː/<üröklü>

Summary: 16

Words generated from transliteration values using Transliteration2Glyph

örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>
örökké: <örökke>

Summary: 1000

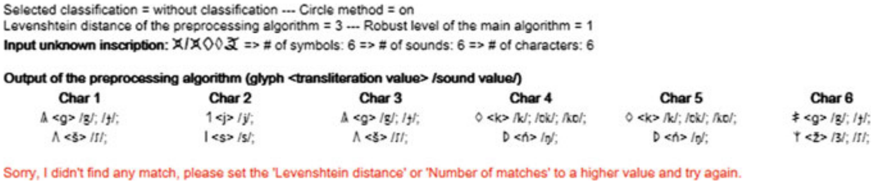
Result	Identified glyph	Transliteration value	Identified word	Hamming distance	Euclidean distance
1		<örükke>	örökké /ørøkkeː/	25.0	8.0
2		<ürökke>	örökké /ørøkkeː/	25.0	8.0
3		<ürükke>	örökké /ørøkkeː/	25.0	8.9
4		<üröklü>	öröklő /ørøkløː/	33.0	11.4
5		<örökke>	örökké /ørøkkeː/	35.0	9.1
6		<ürükke>	örökké /ørøkkeː/	35.0	9.4

Run time of the algorithm: 23.419 sec.

Fig. 8 Result of the SID algorithm with three matches and increased number of matches

Table 6 Input parameters of test-run 6

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	Without classification	On	3	1

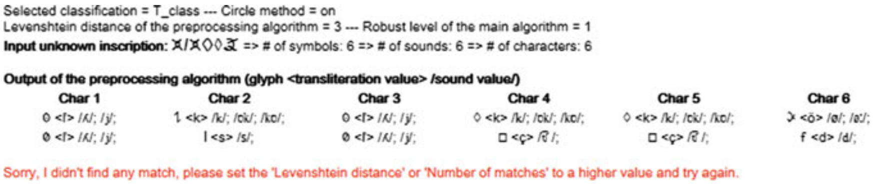


```
Selected classification = without classification --- Circle method = on
Levenshtein distance of the preprocessing algorithm = 3 --- Robust level of the main algorithm = 1
Input unknown inscription: Ӿ/Ӿ◊◊Ʒ => # of symbols: 6 => # of sounds: 6 => # of characters: 6

Output of the preprocessing algorithm (glyph <transliteration value> /sound value/)
      Char 1              Char 2              Char 3              Char 4                  Char 5                  Char 6
  ʌ <g> /g/; /ɟ/;      1 <j> /j/;        ʌ <g> /g/; /ɟ/;   ◊ <k> /k/; /ɒk/; /kɒ/;  ◊ <k> /k/; /ɒk/; /kɒ/;   ‡ <g> /g/; /ɟ/;
   Λ <š> /ʃ/;          | <s> /s/;         Λ <š> /ʃ/;            D <ń> /ɲ/;              D <ń> /ɲ/;          ϒ <ž> /ʒ/; /ʃ/;

Sorry, I didn't find any match, please set the 'Levenshtein distance' or 'Number of matches' to a higher value and try again.
```

Fig. 9 Result of the SID algorithm without any matches

Table 7 Input parameters of test-run 7

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	T-classification	On	3	1



```
Selected classification = T_class --- Circle method = on
Levenshtein distance of the preprocessing algorithm = 3 --- Robust level of the main algorithm = 1
Input unknown inscription: Ӿ/Ӿ◊◊Ʒ => # of symbols: 6 => # of sounds: 6 => # of characters: 6

Output of the preprocessing algorithm (glyph <transliteration value> /sound value/)
      Char 1              Char 2                  Char 3              Char 4                  Char 5                 Char 6
  ϴ <l> /ʎ/; /j/;   ɿ <k> /k/; /ɒk/; /kɒ/;   ϴ <l> /ʎ/; /j/;   ◊ <k> /k/; /ɒk/; /kɒ/;  ◊ <k> /k/; /ɒk/; /kɒ/;  ɔ <ö> /ø/; /øː/;
  ϴ <l> /ʎ/; /j/;        | <s> /s/;         ϴ <l> /ʎ/; /j/;        □ <ç> /ʧ/;              □ <ç> /ʧ/;           f <d> /d/;

Sorry, I didn't find any match, please set the 'Levenshtein distance' or 'Number of matches' to a higher value and try again.
```

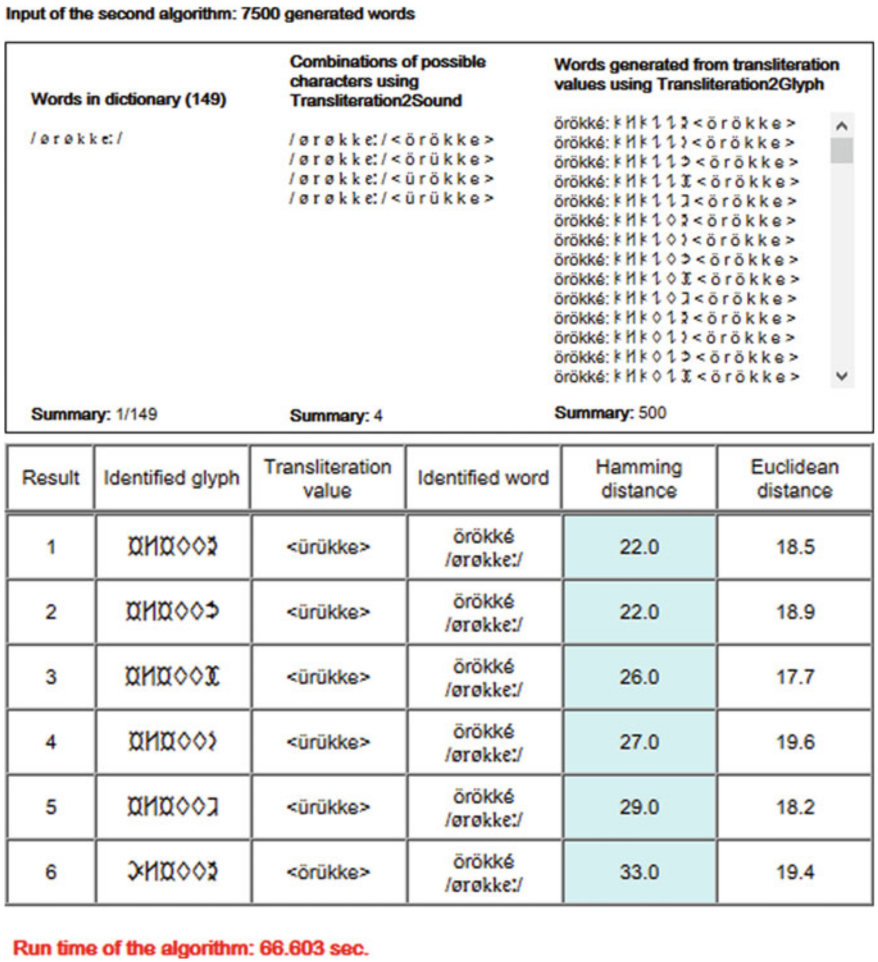
Fig. 10 Result of the SID algorithm without any matches

Table 8 Input parameters of test-run 8

Input settings of the algorithm				
Number of matches	Classification	Circle method	Levenshtein distance	Algorithm robustness
2	V-classification	On	3	1

5 Conclusions

The paper describes a complex procedure called SID to identify the meaning of undeciphered historical inscriptions with using similarity metrics and vector operations based on the topological features of the symbols and glyphs. The SID software uses different methods for accelerating the algorithm and decreasing the processing time and also improving the efficiency of the algorithm and increasing the deciphering probability of unknown inscriptions. After the short introduction of the algorithms and methods, the performance of the system was demonstrated on a real deciphering case. The undeciphered word built from symbols was deciphered successfully by the algorithm. The effects of the implemented improvements were demonstrated by several test-runs with different parameter settings.



Input of the second algorithm: 7500 generated words

Words in dictionary (149)	Combinations of possible characters using Transliteration2Sound	Words generated from transliteration values using Transliteration2Glyph
/ørøkkeː/	/ørøkkeː/<örökke>	örökké: [illegible] <örökke>
	/ørøkkeː/<örükke>	örökké: [illegible] <örökke>
	/ørøkkeː/<ürökke>	örökké: [illegible] <örökke>
	/ørøkkeː/<ürükke>	örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
		örökké: [illegible] <örökke>
Summary: 1/149	Summary: 4	Summary: 500

Result	Identified glyph	Transliteration value	Identified word	Hamming distance	Euclidean distance
1	[illegible]	<ürükke>	örökké /ørøkkeː/	22.0	18.5
2	[illegible]	<ürükke>	örökké /ørøkkeː/	22.0	18.9
3	[illegible]	<ürükke>	örökké /ørøkkeː/	26.0	17.7
4	[illegible]	<ürükke>	örökké /ørøkkeː/	27.0	19.6
5	[illegible]	<ürükke>	örökké /ørøkkeː/	29.0	18.2
6	[illegible]	<örükke>	örökké /ørøkkeː/	33.0	19.4

Run time of the algorithm: 66.603 sec.

Fig. 11 Result of the SID algorithm with one match

The algorithm utilizes the determined characteristic topological features of the glyphs and undeciphered symbols which are the basis of this research. It was found that with the correct definition and selection of the topological features combined with the appropriate similarity metric a very effective deciphering results can be achieved along with using low hardware resources.

The presented results provide an overall picture about the deciphering steps of the examined inscriptions. Our research is not a widely used OCR method as we focused on deciphering ancient hard to read inscriptions. When comparing our algorithm to the existed methods we found distinct solutions for the same deciphering problem. Our approach is different from the researches described in the introduction. The common point in them is that those focus to identify variation of glyphs and find the typical glyph associated to them, that has an interpretation (usually a sound value). In most cases we are facing with hard to read or defective glyphs and sometimes undeciphered glyphs, which has no interpretation (their sound values are unknown a priori). In spite of the difficulties the proposed SID algorithm provides very promising results. Our research is a pioneer in deciphering of historical Székely-Hungarian Rovásh inscriptions with computational palaeography methods. The present study concentrates on the description of Székely-Hungarian Rovásh inscriptions, but the SID algorithm could be extended to other scripts. The introduced approach may give a support for the paleographers in deciphering ancient inscriptions and exploring the relations among historical scripts additionally.

References

1. G. Jäger, Computational historical linguistics. Theor. Linguist. **45**(3–4), 151–182 (2019)
2. T. Rama, J.-M. List, J. Wahle, G. Jäger, Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2018, vol. 2 (Short Papers), pp. 393–400 (2018)
3. A. Chaudhuri, K. Mandaviya, P. Badelia, S.K. Ghosh, in *Optical Character Recognition Systems for Different Languages with Soft Computing*. Studies in Fuzzies and Soft Computing (Springer, 2017)
4. D. Chen, J.-M. Odobez, H. Bourlard, Text detection and recognition in images and video frames. Pattern Recogn. **37**, 595–608 (2004)
5. M. Rahman, S. Islam, R.S. Aktaruzzaman, Convolutional neural networks performance comparison for handwritten Bengali numerals recognition. SN Appl. Sci. **1** (2019) (Article number: 1660)
6. G. Amato, F. Falchi, L. Vadicamo, Visual recognition of ancient inscriptions using convolutional neural network and fisher vector. ACM J. Comput. Cult. Heritage **9**(4) (2016) (Article 21)
7. B. Snyder, R. Barzilay, K. Knight, A statistical model for lost language decipherment, in *Conference: ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 1048–1057 (2010)
8. S. Daggumati, P. Revesz, Data mining ancient scripts to investigate their relationships and origins, in *Conference: The 23rd International Database Applications & Engineering Symposium*, article no. 26, June 2019, pp. 1–10 (2019)
9. G. Hosszú, F. Kovács, Topological analysis of ancient glyphs, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, October 2016
10. P. Rousopoulos, M. Panagopoulos, C. Papaodysseus, F. Panopoulou, D. Arabadjis, S. Tracy, F. Giannopoulos, S. Zannos, A new approach for ancient inscriptions' writer identification, in *17th International Conference on Digital Signal Processing (DSP)*, July 2011
11. Y.-S. Lin, J.-Y. Jiang, S.-J. Lee, A similarity measure for text classification and clustering. IEEE Trans. Knowl. Data Eng. **26**(7), 1575–1590 (2014)
12. A. Barmpoutis, E. Bozia, R.S. Wagman, A novel framework for 3D reconstruction and analysis of ancient inscriptions. Mach. Vis. Appl **21**(6), 989–998 (2010)
13. P. Sapirstein, Segmentation, reconstruction, and visualization of ancient inscriptions in 2.5D. J. Comput. Cult. Heritage **12**(2) (2019) (article 15)
14. G. Hosszú, The Rovas: a special script family of the central and eastern European languages, in *Acta Philologica 44* Wydział Neofilologii Uniwersytet Warszawski. Warszawa, pp. 91–102
15. L.L. Tóth, R.E.I. Pardede, G. Hosszú, Novel algorithmic approach to deciphering Rovash inscriptions, in *Encyclopedia of Information Science and Technology*, 3rd edn., ed. by M. Khosrow-Pour (Information Science Reference, Hershey, PA, 2015), pp. 7222–7233

16. N. Das, J.M. Reddy, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, D.K. Basu, A statistical-topological feature combination for recognition of handwritten numerals. Appl. Soft Comput. **12**(8), 2486–2495 (2012)
17. S. Bag, G. Harit, P. Bhowmick, Topological features for recognizing printed and handwritten Bangla characters, in *MOCR_AND '11: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, article no. 10, pp. 1–7, September 2011
18. H. Tirandaz, M. Ahmadnia, H. Tavakoli, Geometric-topological based arabic character recognition. A New Approach J. Theor. Appl. Inf. Technol. **95**(15), 3692–3702 (2017)
19. R.I. Zaghloul, E.F. AlRawashdeh, D.M.K. Bader, Multilevel classifier in recognition of handwritten Arabic characters. J. Comput. Sci. **7**(4), 512–518 (2011)
20. L.L. Tóth, G. Hosszú, A new topological method for examining historical inscriptions. J. Inf. Technol. Res. **12**(2) (2019)
21. G. Hosszú, *Heritage of Scribes: The Relation of Rovas Scripts to Eurasian Writing Systems,* 3rd edn. (Rovas Foundation, Budapest, 2013)
22. W.D. Basson, M.H. Davel, Category-based phoneme-to-grapheme transliteration, in *Conference: Interspeech* at Lyon, France, August 2013, pp. 1956–1960 (2013)
23. G.S. Shehu, A.M. Ashir, A. Eleyan, Character recognition using correlation & hamming distance, in *23rd Signal Processing and Communications Applications Conference (SIU)*, June 2015
24. J.M. Cunderlik, D.H. Burn, Switching the pooling similarity distances: Mahalanobis for Euclidean. Water Resour. Res. **42**, W03409 (2016)
25. L.L. Tóth, R.E.I. Pardede, G.A. Jeney, F. Kovács, G. Hosszú, Application of the cluster analysis in computational palaeography, in *Handbook of Research on Advanced Computational Techniques for Simulation-Based Engineering,* ed. by P. Samui (Engineering Science Reference, Hershey, PA, 2016), pp. 525–543
26. R.E.I. Pardede, L.L. Tóth, G.A. Jeney, F. Kovács, G. Hosszú, Four-layer grapheme model for computational palaeography. J. Inf. Technol. Res. (JITR) **9**(4), 64–82 (2016)
27. M.E.W. Putra, I. Supriana, Structural offline handwriting character recognition using Levenshtein distance, in *The 5th International Conference on Electrical Engineering and Informatics*, August 2015, pp. 35–40 (2015)
28. C. Zhao, S. Sahni, String correction using the Damerau-Levenshtein distance, in *7th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2017): Bioinformatics*, vol. 20, article number: 277
29. M. Stauffer, P. Maergner, A. Fischer, R. Ingold, K. Riesen, Offline signature verification using structural dynamic time warping, in *Conference: 2019 International Conference on Document Analysis and Recognition (ICDAR)*, September 2019
30. J.K. Kanimozhi, Skeletal graph based topological feature extraction of an object. J. Comput. Appl. **5**(EICA2012–1), 111–118 (2012)

Urban Public Traffic Network Optimization Based on Simulation Technology

Juenuo Yang, Li Wang, and Xiaoning Zhu

Abstract In this paper, we surveyed and counted conventional bus routes within the influence of Shijiazhuang Metro Line 3. The 12 bus routes of main study were determined by the number of relevant stops and the repeated length. We optimized 12 bus routes by means of cancelling the bus stop, extending the route, shortening the route and partially adjusting route to obtain a scheme for the adjustment of the bus network. The road section near the No. 2 middle school station was selected and simulated with VISSIM software. The models before and after adjustment were compared to prove the feasibility of adjustment scheme. Adjusting bus routes within the influence of Metro Line 3 will provide reference for the bus network optimization and help promote the sustainable development of urban traffic.

Keywords Optimization of bus network · Simulation technology · VISSIM · Metro line

1 Introduction

With the development of the economy, the number of urban populations and private cars has increased, which results in increasingly serious traffic problems. In order to solve the problem of urban traffic congestion and meet the needs of public travel, it is necessary to give priority to the development of public transportation in cities. Metro has developed rapidly in recent years due to its advantages of speed and comfort. With the commissioning of the Shijiazhuang Metro, it will definitely bring a certain impact to the regular bus. After the opening of the metro, it is necessary to adjust the bus route network to make the metro and the regular bus develop synergistically.

J. Yang (✉) · L. Wang · X. Zhu
Beijing Jiaotong University, Beijing, China
e-mail: 19125780@bjtu.edu.cn

L. Wang
e-mail: liwang@bjtu.edu.cn

X. Zhu
e-mail: xnzhu@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_31

After the opening of metro, the bus network related to it will be changed. This paper optimizes the related bus routes to improve the service quality of public transportation. Simulate the optimized bus routes and compare them with the pre-optimized routes to see if the solution is feasible. In recent years, traffic simulation technology has developed rapidly, and it has become an important tool for transportation majors and transportation planning and design units. Traffic simulation techniques were used to effectively predict and evaluate the feasibility of improved solution.

2 Literature Review

There are many achievements in bus network optimization. Many experts and scholars have established different models from multiple angles and used a variety of optimization methods to optimize the bus network. Héctor Cancela, Antonio Mauttone and María E. Urquhart proposed a mixed integer linear programming model, which fully considered the traveler's behavior of travelers and quantified it as a time cost [1]. In consideration of the unknown demand, Kun An and Hong K. Lo introduced the concept of bus line service reliability to measure the actual traffic demand, and used the set gradient method and neighborhood search algorithm to solve it [2]. Baoyu Hu, Shumin Feng and Cen Nie defined the competition coefficient and cooperation coefficient between lines. They proposed a method for measuring the relationship between bus line competition and cooperation [3]. Sijia Zhang established the passenger flow allocation model of bus network to study the influence of the average travel distance on the competitiveness of long-distance bus lines [4]. Katerina Vakulenko and Katerina Kuhtin used modeling and simulation methods to estimate the main KPI in transportation, and formed the best bus network in suburbs [5].

Some experts optimized bus network based on the operation of rail transit. Yang Sun, Xiaonian Sun and Qingfeng Kong took the integrated network performance optimization as the research goal, and proposed the optimization and adjustment method of the conventional public transit network based on new rail transit line [6]. Di Huang regarded the planning of connecting routes as a traveling salesman problem and used a two-level planning model to solve the service frequency of each transportation mode [7]. Jian Wang, Danqing Shen and Ziyang Wang proposed the adjustment method of the bus line by analyzing the coordination relationship between the common line part of the rail transit line and the conventional bus line [8]. Ning Song and Xiaowei Shi took Ningbo as an example to study the optimization of conventional bus stops and adjust bus routes based on rail transit stations [9]. Baoqing Li and Yang Sun studied the optimization strategy of the new public transport connecting rail transit line from the perspective of the optimization methods of the bus line network and the coordination of connecting bus operations [10].

3 Feature Comparison of Conventional Bus and Metro

The integrated public transport network includes the conventional bus network and the metro network, which are different in many ways. Conventional bus and metro belong to two different ways of urban public transportation. They are the vehicles chosen by urban residents for travel. The two have different aspects in terms of comfort, safety, on-time rate and convenience. After the official operation of the metro, it will attract some passengers from conventional buses. In order to make up for the shortcomings of the low density of the metro network, it is necessary to optimize the bus network within the scope of its influence. Both metro and conventional public transport need to complement each other to build an integrated, comprehensive transportation system that is efficient, safe and environmentally friendly.

In terms of service capacity, the average passenger load of the subway is about 10 times that of the conventional bus. The carrying capacity of the conventional bus is about 120, and the carrying capacity of the subway is about 1200. In terms of driving speed, the speed of the conventional bus is 15–20 km/h, and the speed of the subway is 35–40 km/h. The comparison of service scope is as follows (Tables 1, 2, 3, 4, 5 and 6).

The subway travels underground. It does not occupy land resources, and it is not affected by problems such as ground traffic congestion. The subway uses electric energy and consumes less energy. However, the construction cost of the subway is high and the construction period is long. The energy comparison of metro and bus is as follows.

After the above analysis, metro and conventional bus have their own advantages and disadvantages. The two should coordinate development each other to optimize the urban traffic network and solve urban traffic problems.

Table 1 Comparison of service scope

Transport mode	Transport distance (km)	Station distance (km)	Service area (km^2)	Reachable city area (km^2)
Bus	10–15	500–800	314–706	314–706
Metro	20–35	>1200	1472–2885	1962–3846

Table 2 Comparison of energy

Transport mode	Energy consumption (kcal/p km)	CO_2 emission (g/p km)	Line Cost (million dollars/km)	Unit dynamic floor area (m^2/p)
Bus	154	19.4	0.9–3.4	1.0
Metro	77	3	17	0

Table 3 Adjustment method

Adjustment method	Detailed explanation	Appropriate types
Extending route	Extend the route and create new stops in extended route	The bus route is close to the subway line, and two stations are connected after the extension
Shortening route	Shorten the route and cancel the stops in shortened route	Some bus routes have high coincidence with the subway line
Cancelling route	Cancel the bus routes or stops	The entire bus route coincides with the subway line
Reserving route	Route orientation and site location remain unchanged	The original route does not affect the subway line
Rebuilding new route	Re-establish complete bus routes	Connect newly developed residential areas and increase bus accessibility in new areas
Local adjustment	Adjust stop location or adjust part of the route	Local routes are affected by subway line

Table 4 Length comparison table

Bus routes	Coincident length with line 3 (m)		Total Line length (m)		Coincidence rate	
	Before	After	Before	After	Before (%)	After (%)
No. 3	1.7	0	14	15.2	12.1	0
No. 8	7.4	7.4	13	13	56.9	56.9
No. 13	5	5	14	14	35.7	35.7
No. 17	7	2.64	11.3	5.5	61.9	48.0
No. 20	5.1	5.1	7.5	7.5	68.0	68.0
No. 22	3	0	11.8	11.8	25.4	0
No. 108	4.3	0	15.6	8.6	27.6	0
No. 110	4.6	4.6	14.2	14.2	32.4	32.4
No. 115	2.1	0	13.3	13.3	15.8	0
No. 117	7.4	7.4	16.3	16.3	45.4	45.4
1 ring 1	5	2.9	21.2	17	23.6	17.1
1 ring 2	5	2.9	21.2	17	23.6	17.1

4 Bus Network Optimization

Passengers will choose the mode of travel according to the distance of the nearby subway station or the bus stop. Taking the subway station as the center and 0.67 km as radius, people can often reach the subway station by foot or by bicycle. This area is also called the subway station’s one attraction range. If there are bus stops and bus lines within this range, the bus stop is called the relevant stop, and the bus

Table 5 Survey data

Street/Road	Direction	Straight traffic flow (vehicle/h)	Left turn traffic flow (vehicle/h)	Right turn traffic flow (vehicle/h)	Total traffic flow (vehicle/h)
Zhonghua north street	From north to south	1638	144	36	1818
Zhonghua north street	From south to north	1452	48	60	1560
Xingkai road	From east to west	72	165	168	405
Xingkai road	From west to east	96	168	72	336

Table 6 Traffic light configuration

Position	Green light duration (s)	Yellow light duration (s)	Red light duration (s)	Total duration (s)
Go straight and turn right at Zhonghua north street	77	3	90	170
Turn left at Zhonghua north street	35	3	132	170
Pass through at Xingkai road	47	3	120	170

route passing through the bus stop is called the relevant route. Find out the bus stops within the attraction range of each metro station, and then find out the bus routes passing through the relevant bus stops to determine which to be adjusted. The basic conditions of the relevant bus line length and the coincidence length of the rail transit line 3 are counted.

4.1 Bus Route Adjustment Purpose

The adjustment of bus routes is to reduce the unreasonable competition between conventional bus and metro, so that the transportation capacity can be fully utilized. And the time and space connection between the two can be improved to achieve the maximum degree. The adjustment of the conventional bus routes should take into account factors such as operating costs, residents' travel and traffic status to achieve the purpose of reducing the travel time of the resident, alleviating traffic, and facilitating residents' travel.

4.2 Bus Route Adjustment Method

There are mainly several methods to adjust bus routes: extending route, shortening or cancelling route, reserving route, creating new route, and local adjustment.

4.3 Bus Route Specific Optimization Scheme

The following routes are adopted the method of local adjustment: Bus No. 3, 22 and 115. The bus No. 3 route and the subway line have a 1.7 km coincidence in Zhonghua South Street. It is adjusted from the intersection of Yuhua Road and Zhonghua Street to the west through sixth middle school and Haiyuetiandi, from Weiming South Street to the south to Wei'an West Road to the east. Bus No. 22 will transfer the Zhonghua Street section to the Weiming South Street section. After crossing the Wei'an West Road and then going to Hongqi Street, it will coincide with the original route. Bus No. 115 is adjusted to go south along Youyi North Street to Heping West Road and then east to the monument.

The following routes are adopted the method of cancelling stops: Bus No. 8, 13 and 20. The length of the coincidence route between bus No. 8 and metro line 3 reaches 7.4 km, and the No. 8 route cancels Zhonghua and Xingkai intersection and the power supply stops. Bus No. 13 cancels Zhonghua and Heping Intersection, Zhonghua and Xingkai intersection and Juranlijia stops. Bus No. 20 cancels Zhonghua and Heping intersection, Hengfeng hotel and power supply stops.

The following routes are adopted the method of shortening routes: Bus No. 17, 108, 1 ring 1 and 1 ring 2. The starting and ending stops of No. 17 route were changed from Xiao'anshe to Jinbolin. And the starting and ending stops of No. 108 were changed from Chencun to Jinbolin.

The following routes are adopted the method of reserving routes: Bus No. 117 and No. 110.

4.4 Overall Comparison Before and After Adjustment

Compare the overall circuit before and after the adjustment.

The total length of these 12 bus lines before adjustment is 173.4 km, and the overlap length with Rail Transit Line 3 is 57.6 km. The overlap rate is 33.2%. After adjustment, the total length is 153.4 km, and the overlapping length with Rail Transit Line 3 is 37.94 km, The coincidence rate is reduced to 24.7%, which Reduces competition between conventional buses and rail transit.

5 Bus Network Optimization Simulation

5.1 Survey Data

This paper analyzes the road traffic conditions along the Shijiazhuang metro line 3, and selects the road sections along the route that have more buses and larger passenger traffic for simulation. Select the road section near the second middle school. First, conduct a field survey of traffic flow, road conditions, and traffic light duration, then get the following data.

The traffic light cycle at the intersection of Zhonghua North Street and Xingkai Road is 170 s. The specific time is shown in the table below.

In the road section of the study, there are three bus stops at the Zhonghua North Street, including the municipal transportation bureau, the Zhonghua and Xingkai intersection and the Shifang Building. The municipal transportation bureau has three stops.

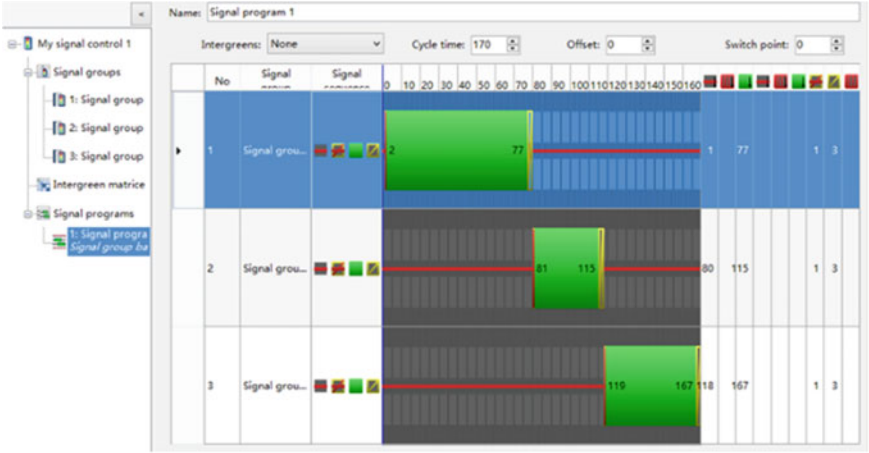
5.2 Build a Simulation Model

First, a view of the area to be studied is intercepted on the map, and the floor plan is imported into VISSIM as a base map. Modify the scaling ratio and position of the basemap in the background parameters. In VISSIM, draw the lanes of Zhonghua North Street, Xingkai Road and Xinhua Road and the link of straight, left turn and right turn. Then input the traffic flow of field investigation into the roads. Set the traffic light at the intersection and edit the signal controller in the signal control option to fix the traffic light at the intersection. The period is 170 s. Finally, create bus stops and bus routes, and set the type, name and length of bus stops (Fig. 1).

5.3 Simulation Result

Configure evaluation indicators and set evaluation parameters and evaluation outputs. Set the queue counter at each intersection of Zhonghua North Street and Xingkai Road, and select the queue length in the evaluation column and simulate. The comparison of queue length before and after adjustment is shown in Table 7.

The VISSIM software was used to simulate the adjusted bus route. The road network and bus stops near the second middle school station were selected. The simulation model was established to compare the road network before and after the adjustment, and the queue length was reduced. The adjusted bus route network can better adapt to the opening of metro line 3.



Name: Signal program 1

Intergreens: None   Cycle time: 170   Offset: 0   Switch point: 0

No	Signal group	Start	End				
1	Signal grou...	2	77	1	77	1	3
2	Signal grou...	81	115	80	115	1	3
3	Signal grou...	119	167	118	167	1	3

Fig. 1 The signal configuration scheme

Table 7 Comparison of queue length before and after adjustment

Position	Average queue length (m)		Maximum queue length (m)	
	Before	After	Before	After
Zhonghua north street (north to south)	64	58	117	110
Zhonghua north street (south to north)	68	60	126	120
Xingkai road (east to west)	16	16	32	30
Xingkai road (west to east)	25	24	46	43

6 Conclusion

In this paper, we surveyed and counted the relevant stations and related lines within the attraction of metro line 3, and optimized bus routes that need to be adjusted by optimization methods such as cancelling the stop, extending the route and shortening the route. The road network near the No. 2 middle school was selected to simulate by VISSIM. The road network parameters before and after bus route adjustment are compared to prove that the optimization scheme is feasible. The innovation is to consider the impact of rail transit operations on bus lines, and combined with simulation to prove the feasibility. This paper has a reference value for the adjustment of the bus line after the operation of metro line. But there are also some limitations. The travel time and cost of passengers are not considered in it.

Acknowledgements This paper was funded by Civil Aviation Administration of China Major projects (grant number: 201501) and the National Key R&D Program of China (grant number: 2018YFB1201403)

References

1. H. Cancela, A. Mauttone, M.E. Urquhart, Mathematical programming formulations for transit network design. Transp. Res. Part B **77**, 17–37 (2015)
2. K. An, H.K. Lo, Two-phase stochastic program for transit network design under demand uncertainty. Transp. Res. Part B Methodol. **84**, 157–181 (2016)
3. B.Y. Hu, S.M. Feng, C. Nie, Bus transport network of Shenyang considering competitive and cooperative relationship. Phys. A **466**, 259–268 (2017)
4. S.J. Zhang, S.P. Jia, B.H. Mao, C.R. Ma, T. Zhang, Influence of passenger trip distance distribution on competitiveness of bus lines in urban rail transit network. J. Zhejiang Univ. (Eng. Sci.) **53**(2), 292–298 (2019)
5. K. Vakulenko, K. Kuhtin, I. Afanasieva, A. Galkin, Designing optimal public bus route networks in a Suburban area. Transp. Res. Procedia **39**, 554–564 (2019)
6. Y. Sun, X.N. Sun, Q.F. Kong, Methodology of bus network optimization and adjustment under operation of new urban rail transit line. J. China Railway Soc **3**, 1–8 (2014)
7. D. Huang, Z.Y. Liu, X. Fu, P.T. Blythe, Multimodal transit network design in a hub-and-spoke network framework. Transp. A Transp. Sci. 706–735 (2018)
8. J. Wang, D.Q. Shen, Z.Y. Wang, Modeling and simulation of coordination relationship between rail and bus lines. J. Syst. Simul. **31**(10), 1995–2009 (2019)
9. N. Song, X.W. Shi, M. Yang, L. Mao, L.X. Shao, A review of research on the status of public transportation lines connecting to Ningbo rail transit network. Spec. Econ. Zone **1**, 53–55 (2020)
10. B.Q. Li, Y. Sun, Optimization strategy of conventional public transport connection based on newly built rail transit line. Transp. World **3**, 11–12+15 (2020)

Special Purpose Network Simulator for Transport Protocol Analysis

Dávid Tegze, Ferenc Kovács, and Gábor Hosszú

Abstract This paper presents a tool for evaluating transport layer congestion control schemes, including various TCP variants. In real computer networks, it is difficult to obtain detailed information about the internal state and operation of network protocols without disturbing protocol behavior. This is especially true for transport layer protocols where timing plays an important role. For deep analysis of protocol mechanisms, network simulation plays an essential role. This paper demonstrates a network simulator tool channel *SimCast.* This simulator focuses on the congestion control methods of the transport layer. Using this special purpose simulator, the effects of network topology and loss models can be investigated in detail. This paper presents a comparison of simulation results from SimCast to the output of other network simulators and real network measurements.

Keywords Transport protocol · Network simulator · TCP · Protocol fairness

1 Introduction

An important means for performance analysis of computer networks is simulation [1]. The most significant advantage of simulation-based evaluation methods compared to real network traffic measurement is the access to the detailed internal state of the studied network and protocols at various levels. This is not possible with real network measurements without disturbing the network and behavior.

D. Tegze (✉) · G. Hosszú
Department of Electron Devices, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: tegze@eet.bme.hu

G. Hosszú
e-mail: hosszu@eet.bme.hu

F. Kovács
Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
e-mail: kovacsf@itk.ppke.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_32

In simulated environment, we can access information that is not available on the endpoints. With the detailed data that is available inside the simulation a direct comparison and evaluation of protocol behavior is feasible.

Simulation has access to network wide information like queue saturations, packet loss ratios of flows etc. Although a large number of packet level simulator is available today [2, 3], we have implemented our own simulation software for reasons explained in the followings.

The special purpose simulator *SimCast* implements the layers of the TCP/IP protocol stack with emphasis on the transport layer and it has a detailed model of congestion control. The SimCast simulator and real network measurement methods together constitute a comprehensive transport protocol analysis system [4].

Demonstration and qualitative validation of the developed simulation is presented in this paper. Moreover, a comparison of different transport level congestion control schemes, including variants of the TCP is presented as well. Protocol mechanisms implemented in various protocols are hard to investigate in a uniform manner; therefore, the simulator SimCast is developed for traffic analysis of network streams. In this article, the basic TCP protocol mechanisms are compared using the SimCast simulator [4]. Our simulation results are presented through examples.

Due to spreading of traffic lacking end-to-end congestion control, congestion collapse may arise in the Internet [5]. This form of congestion collapse is caused by congested links that are sending packets to be dropped later in the network. The essential factor behind this form of congestion collapse is the absence of end-to-end feedback. On one hand, an unresponsive flow fails to reduce its offered load at a router in response to an increased packet drop rate and on the other hand, a disproportionate-bandwidth flow uses considerably more bandwidth than other flows in time of congestion. In order to achieve accurate traffic simulation—being not so TCP-friendly yet—effects of the TCP flow control should be determined [6]. Since there are many kinds of TCP and other transport protocol implementations with various flow control mechanisms, this investigation can be rather difficult [7, 8].

Up to now many comparisons have been done [8]. For example, Wang et al. reviewed the TCP-friendly congestion control schemes in the Internet [9]. They differentiated two groups of TCP-friendly congestion control algorithms as follows: (1) end-to-end and (2) hop-by-hop congestion control mechanisms. The end-to-end mechanisms are grouped into (i) AIMD-based schemes (AIMD: Additive Increase Multiplicative Decrease) with the window- and rate-adaptation schemes; (ii) modeling-based schemes, including equation-based congestion control schemes and the so-called model-based congestion schemes; and (iii) a combination of AIMD-based and modeling-based mechanism. Wang's classification is mostly used in present discussion, too.

This paper demonstrates a novel simulator for transport protocols and its results are collated with measurement data and other simulators.

Finally, conclusions are drawn and work to be done identified.

2 Properties of Existing Network Simulators

Before presenting the developed simulation method this section demonstrates the basic properties of the well-known network simulators existing today—the free ns-2 and ns-3 software and the commercial OPNET simulator [10, 11]. These simulators implement discrete event simulation (DES) [7]. DES keeps track of a list of system events ordered by time and the value of state variables is updated at the occurrence of these events [12]. The simulation time is incremented according to the list of processed events [13].

The simulator *ns-2* is the open source descendent of the REAL simulator [14–17] supported by the DARPA, numerous universities and companies [18]. The ns-2 is written in C++ and object-oriented TCL languages mainly on Linux and UNIX operating systems. The ns-2 suite includes a TCL language-based animation tool for the visualization of simulation results. The ns-3 simulator is not the result of ns-2 evolution. It is completely rewritten in C++ language and has python bindings for scripting and visualization.

The OPNET simulator is also an object-oriented simulation tool that implements a hierarchical model structure. This allows the application and development of models with different abstraction levels. Different detail levels are possible. The simulation process supports node-based and network-based modeling. The OPNET has a very comprehensive model library and it allows the development of custom model components using the C++ language. The user has the possibility to provide a state machine to describe protocol behavior. The OPNET has a menu driven graphical user interface to create the simulation setup and it has its own graphing and animation tool for the analysis of results.

The *SimCast* simulator implements discrete, dynamic and stochastic simulation method. It is discrete since the state variables can only change at events occurring at discrete times. It is dynamic since the SimCast traces the changes of state variables in time and it is stochastic since the simulation setup can contain stochastic components like the various loss models. Of course, it is possible to create completely deterministic systems when all the simulated model components in the setup of the SimCast are deterministic.

Unlike the previously mentioned simulators, SimCast uses fixed time steps. The simulation time is updated with fixed increments. All event that occurred during a simulation time step is treated as if they occurred at the end of the step. This method has its benefits and drawbacks. On one hand, simulating the effects of events at the end of the time slice introduces some error. This error can be decreased by using smaller time steps; however, this will increase the required computation power. On the other hand, the main advantage of this kind of simulation is synchronism. This means that this method allows the distributed parallel execution [19].

The topology model of SimCast allows the interconnection of two or more SimCast instances over a real network. In such setup, the SimCast simulator instances simulate subtopologies. This kind of distributed operation requires the synchronous execution that is provided by the time step-based simulation architecture. This setup allows the execution of simulated flows on a real network link which is shared with real flows.

Evaluation of transport protocol properties is carried out using the SimCast, which is focused on the simulation of the transport layer. In SimCast, flow attributes are available at different layers of the network stack. For example, flow related data can be associated to individual transmitted packets. Combining data at different levels and nodes allows the detailed analysis of network traffic and taking information of all network layers into consideration in the same time. In SimCast, emphasis is put to the analysis of transport protocols.

In order to validate SimCast, our simulation results were both compared against real network measurements and simulation results [20]. Real network measurements were carried out using the *Tcpprobe* software tool. With Tcpprobe it is possible to access internal protocol variables of the transport layer. By evaluating, the time series of protocol attributes in the simulations and measurements, the behavior of the simulator SimCast was verified (see details below). Additionally, for further validation the operation of the SimCast simulator was compared with the ns-3 network simulator (see details below).

To set up a transport protocol test environment configuration being relevant to congestion control can be specified in detail while less relevant attributes have a simplified configuration. To have a functional transport protocol evaluation environment network topology, queuing and flow attributes should be defined, together with various transport layer properties. SimCast can record transport layer information in detail and it is possible to use the simulator's data visualizer to extract and demonstrate the internal state of concurrent flows. It is possible to register custom protocol information to the recording and visualization system. These custom attributes can then be analyzed together with the standard properties and compared against each other. Custom loss models are supported to simulate transient and steady state behavior. The time window for loss definitions is configurable and several definitions are allowed. In case of overlapping loss definitions, the union of all generated losses is applied.

The SimCast can calculate various fairness metrics (see the description below). These metrics can be accessed from different levels of the protocol stack. In the simulation environment, it is possible to influence congestion control algorithms based on the perceived fairness metrics [21]. The SimCast supports batch executions since functional components of the simulator software are well separated from the control and visualization. The graphical user interface, the manual and batch execution and evaluation component are loosely coupled. The SimCast has the possibility to execute multi-dimensional simulations by varying several protocol parameters in nested executions. In such simulation setups, it is possible to evaluate the effect of changing one protocol attribute and leaving all other parameters unchanged. Using such parametrization, it is possible to extract a "cutaway view" of the multi-dimensional result set.

The SimCast has a graphical topology editor and execution control interface to manage simulations. Additionally, an offline evaluation tool is developed for SimCast. Using this evaluation tool, it is possible to analyze various transport protocol properties from the simulation data sets. To facilitate evaluation of batch simulations

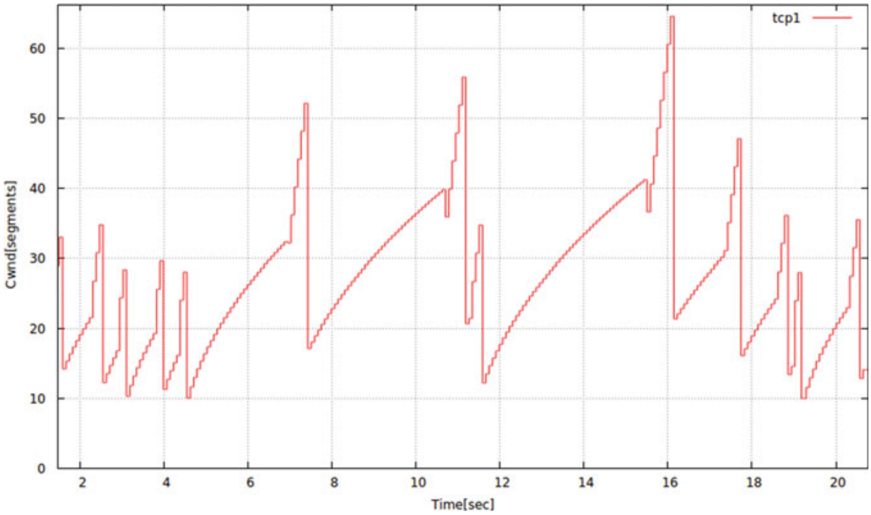


Fig. 1 CWND trace from the SimCast simulator

several scripts are available. These scripts extract relevant information from batch data and provide visualization possibilities.

In order to compare and validate the SimCast simulator real network measurements in Linux environment have been carried out. The Linux kernel allows high precision recording of internal kernel variables using the *kprobe* infrastructure [22]. *Tcpprobe* is using this method to record time series of TCP state variables. To generate network traffic for the validation of Simcast the *iperf* tool was used. Iperf allows the creation of communication endpoints and sending bulk data flows between them [23]. For testing the response of real TCP flows to loss events, the *traffic control* (*tc*) Linux tool was used to set sending rate limitations and shaping bandwidth of real network flows [24]. Congestion window (*Cwnd)* [25] traces under similar conditions are presented for simulated (Fig. 1) and measured (Fig. 2) flows.

In the reference ns-3 simulation results, Fig. 3 shows the Cwnd and Fig. 4 plots the RTT for a TCP Reno flow exposed to random packet losses. Comparing the Cwnd plots of SimCast, tcpprobe and ns-3 it is apparent that these plots are qualitatively similar. Our comparison is demonstrated in [20] in more details.

3 SimCast Validation with ns-2

The goal of the validation efforts with SimCast is to decide whether the SimCast models the real system accurately with respect to the objective pursued. A model can only approximate the real system; therefore, our validation checks its correctness with respect to the relevant aspects only. The validation has been carried out by

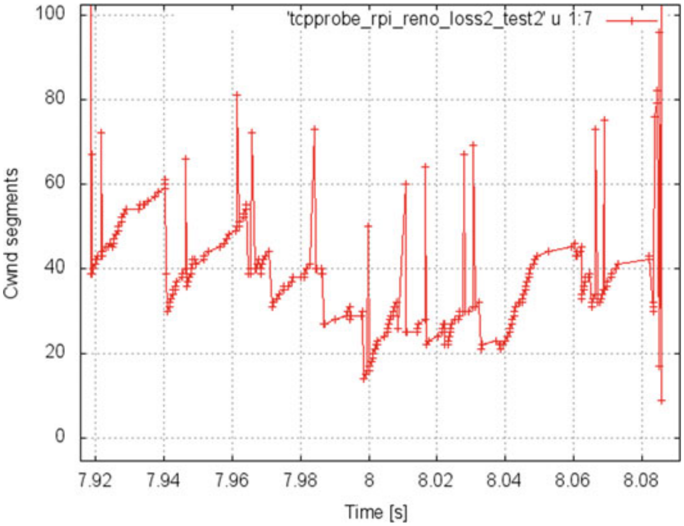


Fig. 2 CWND trace from Tcpprobe over Ethernet network

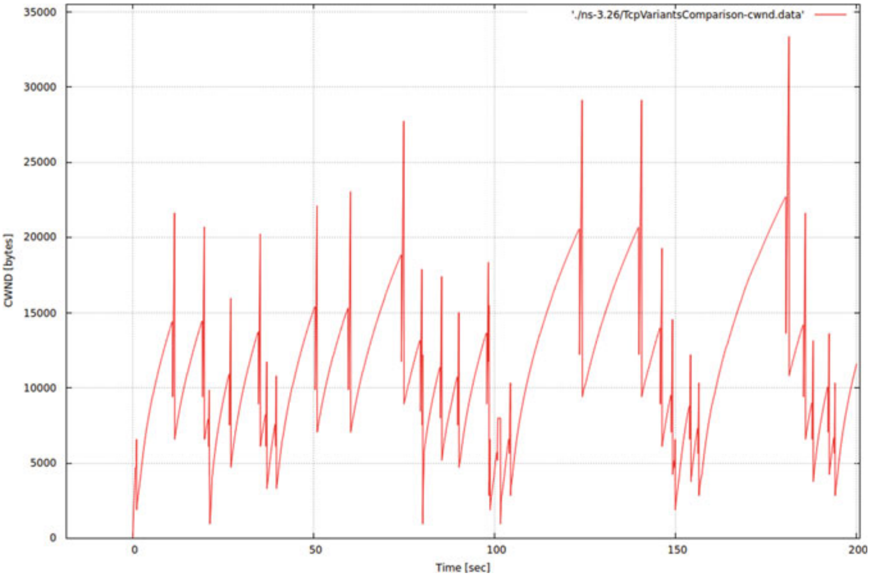


Fig. 3 CWND trace from the ns-3 simulator

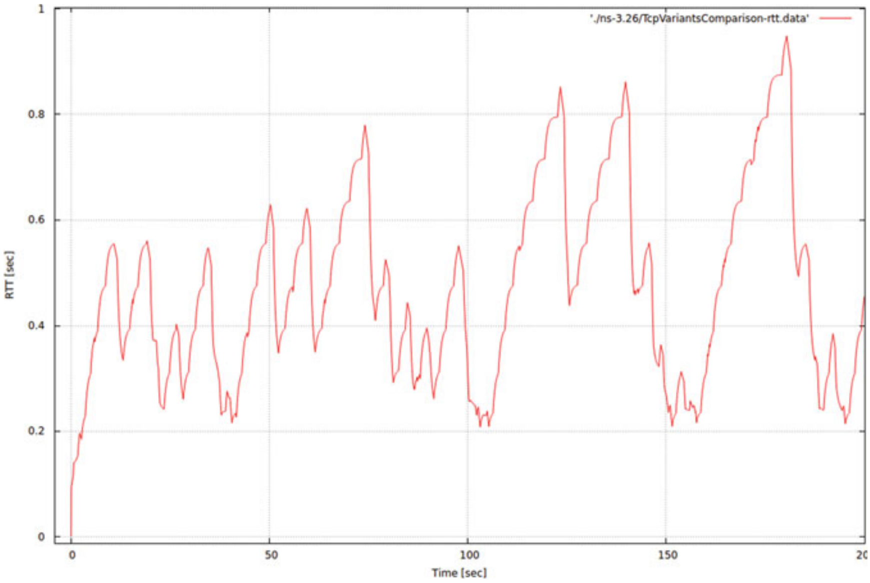


Fig. 4 RTT estimation trace from the ns-3 simulator

executing simple simulations and comparing the results with measurement data and with the output of other valid simulators.

Validation plays an important role in different simulation techniques. The lack of validation can cause the incorrect implementation of simulation models that can lead to incorrect conclusions.

In the followings, different TCP based congestion control methods are investigated. The results of SimCast simulations are compared with simulation outputs from ns-2 simulations. [26, 27].

During the validation process, the mentioned ns-2 results are used as reference. The SimCast based setup is very similar to the ns-2 configuration nonetheless reproducing the exact same configuration is not possible because of the differences in parametrization of two simulators. This allows a qualitative comparison of the two tools. The presented figures allow us to carry out direct comparisons between the well-known ns-2 and the SimCast simulators. They give the possibility of a direct comparison of the reaction to packet loss events and various network conditions. In the followings, the output of SimCast is compared to results from the ns-2. This allows the validation of transport layer congestion control methods implemented in the SimCast simulator.

The first step of the validation examines a Reno TCP protocol entity. Simulation results are shown on Figs. 5a and 6a. The goal of the simulation is the detailed analysis of sending and receiving times in Reno TCP for both simulators. The figures show the evolution of sequence numbers of TCP data segments in time. Since sequence numbers increase monotonously and our goal is the detailed investigation of sending

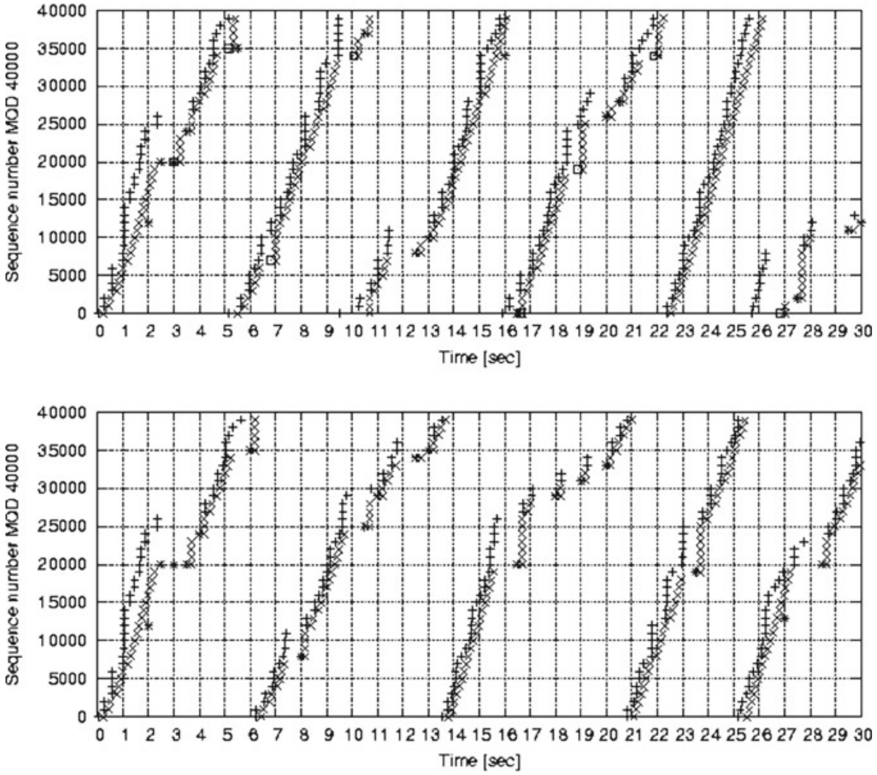


Fig. 5 Tracing sending (+), arrival (x), and retransmission events (* —timeout based retransmission and □ —fast retransmit event) in *SimCast* under the same packet loss rate **a** TCP *Reno* (upper figure) **b** TCP *Tahoe* (bottom figure)

and receiving times, it is practical to apply modulo division on the sequence numbers in order to be able to visualize the sending and receiving patterns in the required detail level. In our examples, the modulo divider was set to 40,000. In the demonstrated simulations, the maximum segment size was set to 1000 bytes. This means that each increasing section in the figure shows the delivery of 40 individual TCP segment.

At $t = 0...1$ s time interval, the *slow start* phase of TCP can be observed. Figure 6a shows the staged increase of Cwnd function, which is caused by the bursty acknowledgement segment arrivals. The ACK clocking mechanism causes the Cwnd to be recalculated on acknowledgement arrivals. Figure 5a shows that the number of transmitted data segments are duplicated in each round-trip time during this period. Figure 7 presents that the TCP Reno protocol entity of ns-2 increases the Cwnd the exact same way during the *slow start* phase. The sharp edges of Cwnd plot increase during the first few round trip times is caused by the unsaturated network buffers. As FIFOs get saturated the sharp edges became smoother and the sending rate increase will become somewhat lower than the initial squared pace observed at

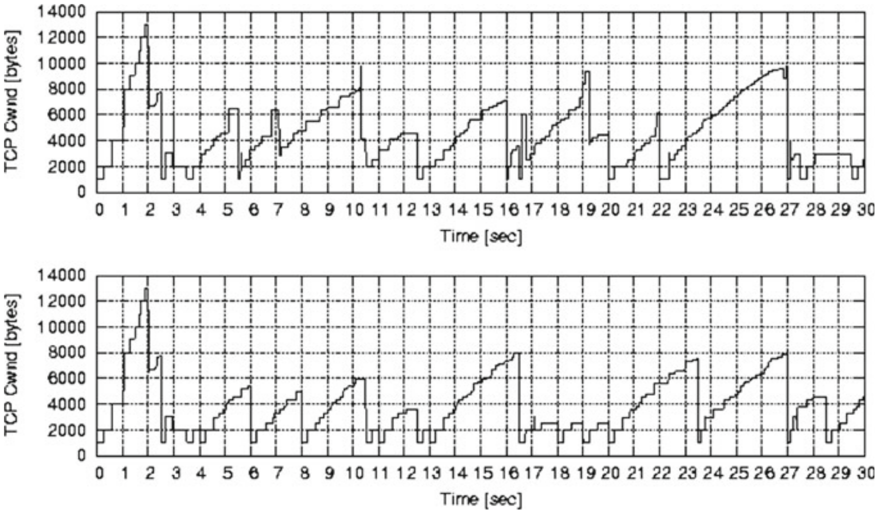


Fig. 6 *Cwnd* trace in SimCast *with t*he same packet loss model **a** TCP *Reno* (upper figure) **b** TCP *Tahoe* (bottom figure)

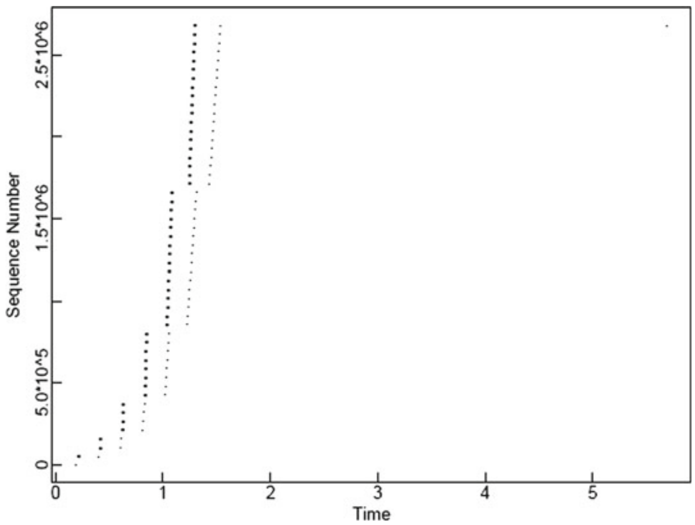


Fig. 7 Trace of *Reno FullTCP* during *Slow Start* [27]

the very beginning. This deceleration is caused by the additional delay caused by the saturated buffers along the communication path. Packets spend more and more time in FIFOs waiting for transmission. Intermediate buffers smooth out sending rate as the network pipe gets filled.

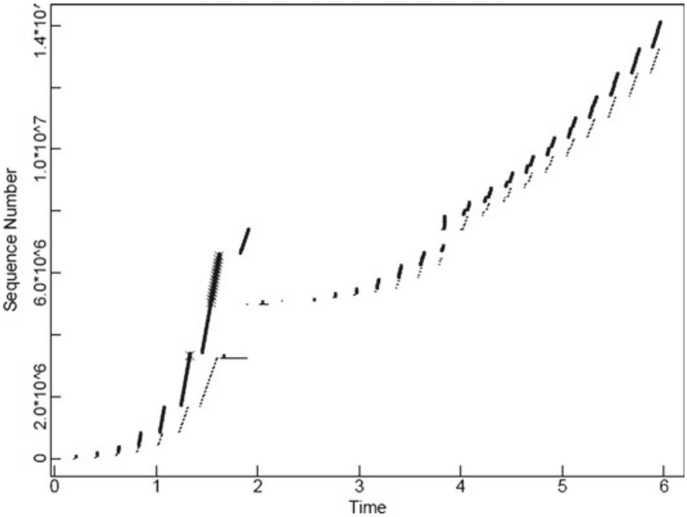


Fig. 8 Trace of *Reno FullTCP* in *Slow Start,* and then in *Congestion Avoidance* state [27]

At certain times, vertical patterns can be seen in the arrival plot. In these cases, the TCP entity passes more than one segments to the application layer. This can occur in situations when a retransmitted segment is successfully delivered, and the continuity of the received buffer is restored. This allows the delivery of several segment data to the application layer. Such situation can be seen on Fig. 5a at t = 7.0. On Fig. 6a this vertical increase of Cwnd is observable at several time intervals (e.g. $t = 7 \ldots 10$ s).

In the ns-2 results Fig. 8 is showing such event at $t = 4 \ldots 6$ s. During this period, the TCP sender stays in the *congestion avoidance* state [27, 28]. In this congestion control state, the *Cwnd* is a sublinear function of time. This sublinear growth can be explained by the extra delay caused by the saturating buffers that slows down the ACK clocking mechanism.

Analyzing Fig. 5a timeout-based packet loss detections can be identified (e.g. $t = 12.5$ s). On Fig. 9 similar timeout events can be observed in the ns-2 TCP Reno simulation plot ($t = 2.4 \ldots 4.6$ s). This loss detection method is required in cases when ACK segments are delayed too much or lost and the *fast recovery* and *fast retransmit* methods cannot detect packet losses and it is not possible to restore the transmitted data stream using these quick recovery methods.

When the flow of ACK segments is ensured, TCP Reno protocol entities can detect and repair by using the significantly more efficient *fast recovery* and *fast retransmission* methods as it is shown by Fig. 5a at $t = 7.0$ s. When these quick methods are used, the network pipe is not depleted, and it is not necessary to switch back to the *slow start* phase. After executing these fast repair methods, the sender can keep sending in *congestion avoidance* mode starting with halved Cwnd. The

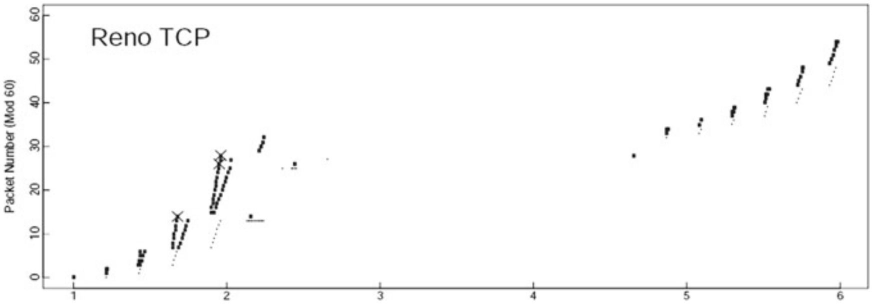


Fig. 9 Trace of *Reno TCP* in the *ns2 network simulator* in case of three consecutive packet loss event; timeout-based recovery [26]

main goal of slow start is to fill the network pipe in order to ensure the best possible network utilization.

From ns-2 plots Fig. 10 shows the triggering of *fast retransmission* phases. These events can be observed as one-time packet transmissions occurring at most once per RTT.

Comparing the simulation results carried out executing *TCP Reno* (Fig. 5a) and *TCP Tahoe* (Fig. 5b) it is visible that the Reno algorithm performed better than Tahoe during the examined 30 s period considering effective throughput. Summing the monotonously increasing sections of the arrival plot the effective throughput is estimated to be $5 * 40{,}000 + 12{,}000 = 212{,}000$ bytes while this measure is $5*40{,}000 = 200{,}000$ bytes in case of Tahoe. Effective throughput is the successfully delivered data at the receiver. This difference in favor of Reno is caused by the fast recovery and fast retransmission mechanism. These algorithms are included in Reno but not in Tahoe. These mechanisms keep the network pipes from getting empty and making network interfaces idle. Thus, the mentioned two methods improve both throughput and network utilization.

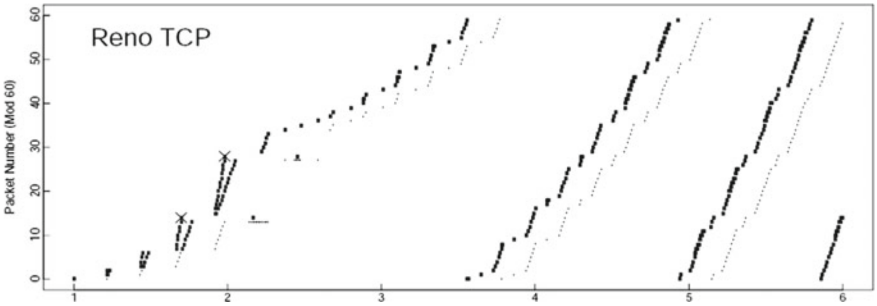


Fig. 10 Fall and Floyd [26]: Trace of *Reno TCP* in the *ns2* network simulator: *fast retransmission-based* recovery

The mentioned methods allow earlier detection and reparation of packet losses that again improves these protocol measures.

At $t = 2.5 \ldots 3.5$ s on the Tahoe plot it is visible that the sender is suffering. The success rate for the sender to deliver data segments is poor. This problem is caused by the continuously saturated queues along the communication path. On the Tahoe figure (Fig. 6b) it is visible that it took three timeout-based retransmissions for the sender to recover the situation. The mentioned figures show that Reno provides a significantly better performance in this respect, too. Due to the triple ACK detection mechanism, Fast retransmit can detect non-series packed losses quickly.

4 Main Features of the Simulator SimCast

The developed simulator software follows object-oriented design principles, which results in modular structure. It uses a non-preemptive scheduler for the accurate execution of simulated objects. The scheduling is multi layered; these layers correspond to the simulated protocol layers. Most scheduling properties can be set on a layer basis. This scheduling system is capable of the execution of several protocol entities concurrently. The simulator is able to handle complex topologies and detailed network node configurations, especially for the transport layer.

The implemented protocol entities are realized as deterministic finite-state-machines. Some of the simulations carried out by SimCast were fully deterministic. However, applying non-deterministic loss models, the simulations result in non-deterministic output, of course. SimCast has both deterministic and non-deterministic loss models built in. They can be selected arbitrarily at configuration time.

The developed network simulation method implements the detailed model of transport layer congestion control. The SimCast simulator implements several transport layer protocols and protocol mechanisms. Special purpose transport protocols like the TCPLIKE and TFRC is supported together with the different variants of TCP. From the TCP protocol family Tahoe, Reno, New Reno and Vegas is implemented. Certain protocol options are also implemented that can be configured for the above protocols like SACK or Limited transmit. Certain protocol functions have more than one slightly different implementations. These can be selected during configuration time.

The SimCast support the calculation of important congestion control related protocol measures. SimCast can execute online and offline evaluation for utilization, smoothness, aggressiveness and fairness. Various bandwidth allocation fairness measures are supported such as Variance, coefficient of variation (CoV), min to max ratio, normalized distance to optimum, Raj Jain's fairness index [29].

5 Conclusions

In this paper we presented a validation study of the SimCast packet level network simulator. Our study analyzed the sending and receiving patterns of two different TCP congestion control protocols. As the presented results show, packet level traces are available which allow the investigation of complex transient behavior of transport protocols. A detailed comparison was carried out on the SimCast and the well-known ns-2 network simulators. Besides this, the study executed a comparison between the SimCast TCP simulation results and real network TCP flow measurements. These measurements are based on the kprobe network protocol tracing infrastructure. These comparisons confirm a good qualitative agreement of the SimCast output both with ns-2 based simulation results and equivalent real network flows. Therefore, we conclude that the developed simulator is suitable for the detailed analysis of transport protocols.

References

1. A.M. Ali, S. Kadry, Performance evaluation of TCP congestion control algorithms using a network simulator, in *Automatic Control, Mechatronics and Industrial Engineering: Proceedings of the International Conference on Automatic Control, Mechatronics and Industrial Engineering (ACMIE 2018),* 29–31 October, 2018, Suzhou, China (CRC Press, March 2019), p. 317
2. B. Liu, Y. Guo, J. Kurose, D. Towsley, W. Gong, Fluid simulation of large scale networks: issues and tradeoff, in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA '99,* Las Vegas, NV, USA, June 1999, pp. 2136–2142 (1999)
3. B. Liu, D.R. Figueirido, Y. Guo, J. Kurose, D. Towsley, A study of networks simulation efficiency: fluid simulation vs. packet-level simulation, in *Proceedings of IEEE INFOCOM 2001,* Anchorage, AK, USA, vol. 3 (2001), pp. 1244–1253
4. M. Orosz, D. Tegze, The SiMCast multicast simulator. in*International Workshop on Control & Information Technology, IWCIT'01,* Ostrava, 19–20 September 2001, pp. 66–71
5. S. Floyd, K. Fall, Promoting the use of end-to-end congestion control in the Internet. IEEE/ACM Trans. Netw **7**(4), 458–472 (1999)
6. J.B. Postel (ed.), Transmission control protocol, DARPA internet program, protocol specification. RFC 793, September 1981
7. Q. He, C. Dovrolis, M. Ammar, On the predictability of large transfer TCP throughput. ACM SIGCOMM Comput. Commun. Rev. (2005)
8. T. Chowdhury. M.J. Alam, Performance evaluation of TCP Vegas over TCP Reno and TCP NewReno over TCP Reno. JOIV: Int. J. Inform. Vis. 3(3), 275–282 (2019)
9. Q. Wang, TCP-friendly congestion control schemes in the Internet, in *Proceedings of the 2001 International Conference on Information Technology and Information Networks (ICII'2001),* Beijing, China, Oct.–Nov. 2001*,* vol. B, pp. 205–210 (2001)
10. X. Chang, Network simulations with OPNET, in *Proceedings of WSC '99,* vol. 1, ed. by P.A. Farrington, H.B. Nembhard, D.T. Sturrock, G.W. Evans (IEEE, Piscataway, NJ, USA, 1999), pp. 307–314
11. L. Campanile, M. Gribaudo, M. Iacono, F. Marulli, M. Mastroianni, Computer network simulation with ns-3: a systematic literature review. Electronics **9**(2), 272 (2020)

12. E.J. Dudewicz, Z.A. Karian, in *Modern Design and Analysis of Discrete-Event Simulations* (IEEE Computer Society Press, Washington, DC, 1985)
13. M.C. Jeruchim, P. Balaban, K.S. Shanmugan, in *Simulation of Communication Systems*, 2nd edn. (Kluwer, New York, 2000)
14. S. Keshav, *Real: a network simulator.* Technical report (University of California at Berkeley, Berkeley, CA, 1988)
15. E. Weingartner, H. vom Lehn, K. Wehrle, A performance comparison of recent network simulators, in *2009 IEEE International Conference on Communications*, Dresden, pp. 1–5 (2009)
16. K. Bazi, B. Nassereddine, Comparative analysis of TCP congestion control mechanisms, in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security (NISS2020)* (Association for Computing Machinery, New York, NY, USA, 2020), Article 13, pp. 1–4
17. L. Breslau, Advances in network simulation. IEEE Comput. (2000)
18. L. Leemis, Input modeling techniques for discrete-event simulations, in *2001 Winter Simulation Conference,* Arlington, VA, vol. 1, pp. 62–73 (2001)
19. S. Bhatt, R. Fujimoto, A. Ogielski, K. Perumalla, Parallel simulation techniques for large-scale networks. IEEE Commun. Mag. **36**(8), 42–47 (1998)
20. G. Hosszú, D. Tegze, F. Kovács, Simulation based comparison of different TCP and TCP-friendly protocols, in *Encyclopedia of Multimedia Technology and Networking*, ed. by M. Pagani, vol. III, 2nd edn. (Information Science Reference, Hershey, PA, 2009), pp. 1307–1315
21. D. Tegze, G. Hosszú, The transport-level requirements of the internet-based streaming, in *Encyclopedia of Information Communication Technology*, ed. by A. Cartelli, M.*,* 2nd edn. (Information Science Reference, Palma, Hershey, PA, 2008), pp. 775–781
22. J. Sun, Z. Li, X. Zhang, Q. He, H. Wang, The study of data collecting based on Kprobe, in *2011 Fourth International Symposium on Computational Intelligence and Design,* Hangzhou, pp. 35–38 (2011)
23. S.S. Kolahi, S. Narayan, D.D.T. Nguyen, Y. Sunarto, Performance monitoring of various network traffic generators, in *2011 UkSim 13th International Conference on Computer Modelling and Simulation,* Cambridge, pp. 501–506 (2011)
24. J. Vila-Carbo, J. Tur-Masanet, E. Hernandez-Orallo, An evaluation of switched Ethernet and Linux traffic control for real-time transmission, in *2008 IEEE International Conference on Emerging Technologies and Factory Automation*, Hamburg, pp. 400–407 (2008)
25. A. Afanasyev, N. Tilley, P. Reiher, L. Kleinrock, Host-to-host congestion control for TCP. IEEE Commun. Surv. Tutor. **12**(3), 304–342 (2010)
26. K. Fall, S. Floyd, Simulation-based Comparisons of Tahoe, Reno and Sack TCP. ACM Comput Commun Rev **26**(3), 5–21 (1996)
27. K. Fall, S. Floyd, T. Henderson, Ns Simulator Tests for Reno FullTCP, July 1997. ftp://ftp.ee.lbl.gov/papers/fulltcp.ps
28. R. Jain, A delay based approach for congestion avoidance in interconnected heterogeneous computer networks. ACM SIGCOMM Comput. Commun. Rev. 56–71 (1989)
29. R. Jain, D. Chiu, W. Hawe, in *A quantitative measure of fairness and discrimination for resource allocation in shared computer systems.* DEC Research Report TR-301 (Digital Equipment Corporation, Hudson, MA, 1984)

Financing Schemes in Supply Chains with a Capital-Constrained Supplier: Coopetition and Risk

Danqin Yang, Jiani Ding, and Nan Yan

Abstract In this paper, we consider a small capital-constrained supplier that can finance from a bank or a peer supplier or a downstream manufacturer in a supply chain. We use a game-theoretical model to analyze different financing schemes. Through analytical comparison, we find that the small supplier will always choose purchase order financing when the production cost is low. As the production cost increases, the internal financing (buyer direct financing and peer financing) dominates external financing.

Keywords Supply chain finance · Game theory · Buyer direct financing · Purchase order financing · Peering financing

1 Introduction

Sourcing from suppliers is common for manufacturers to reduce production cost and it is usual for manufacturers to procure from more than one supplier. Moreover, big manufacturers sometimes purchase from small suppliers who may have inherent unreliable, i.e., performance risk. We consider a small supplier with performance risk as well as capital constraint in our paper.

Trapped by capital constraint is a common dilemma in supply chains, especially for small and medium-sized enterprises (SMEs) or startups. A research reported by the Small Business Credit Survey (SBCS) revealed that among the surveyed SMEs, 53% experienced a financing shortfall when seeking new financing, i.e., the loans obtained cannot yet cover their fund demand, 62% had funding problems, and 37% had trouble meeting daily operational cost. The report also showed that companies

D. Yang (✉) · J. Ding · N. Yan
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China
e-mail: yangdanqin@163.com

J. Ding
e-mail: 783513693@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_33

with weak credit, weak profitability and at their early-stage are particularly short of capital. Most of the companies faces funding shortfalls of $100–$250 K [1].

To mitigate the financial stress and cover the shortfalls of cash flow, companies can borrow from either external parties or internal parties in a supply chain. The external financing includes modes such as bank credit financing and *purchase order financing* (POF) and the internal financing includes modes such as trade credit financing and *buyer direct financing* (BDF). We build up a stylized supply chain model to capture the interactions among two suppliers and one manufacturer in three different financing scenarios (POF, BDF and *peer financing* (PF)).

Our paper contributes to the literature on supply chain financing in the following aspects. First, we analyze financing schemes in a coopetition supply chain structure. We find that the capital-constrained supplier always chooses external financing when the production cost is low; otherwise, the capital-constrained supplier prefers internal financing. Second, we consider the small supplier's inherent unreliability can be improved by exerting costly effort. It is demonstrated that the effort level is only affected by the bank's interest rate.

The rest of this paper is organized as follows. In the next section, we review the most relevant literature. Model settings are described in Sect. 3. Sections 4, 5 and 6 investigate purchase order financing, buyer direct financing, and peer financing respectively. Section 7 addresses the optimal financing scheme for the small supplier. Finally, we conclude the paper and discuss the future research.

2 Literature Review

Our work is closely related to supply chain finance, the majority of which focus on bank credit financing, trade credit financing, buyer direct financing and so on.

References [2–6] focused on comparing different financing schemes and studied the effect of different financing schemes on alleviating the double marginalization and coordinating the supply chain. Tunca and Zhu [7] studied BDF and verified the financing scheme can help to improve the performance of the supply chain. Reindorp et al. [8] investigated POF, and they found that POF can help to mitigate the supplier's financial stress. Tang et al. [9] compared BDF with POF and showed that when a manufacturer has more information about suppliers than a bank, BDF will be a more attractive financing scheme.

Alan and Gaur [10] studied the implications of asset-based lending and demonstrated how managers should make operational and financial decisions when interacting with asset-based lenders. Katehakis et al. [11] investigated the optimal sourcing decision through bank credit financing. Kouvelis and Zhao [12] considered a supply chain composing with two financial constrained parties (both the supplier and the manufacturer) and studied the impact of their different credits on financing decisions.

Our paper is closely related to [9]. The supply chain in [9] comprise with one supplier and one manufacturer. However, we extend the model to consider the case where one manufacturer source from two competitive suppliers. One is a small supplier who has performance risk and capital constraint, and the other is a peer supplier with sufficient capital who can lend money to the small one, i.e., peering financing. Specially, we capture the coopetition feature between the two suppliers in the peer financing scenario. To the best of our knowledge, there is almost no paper considering coopetition in supply chain financing and it is our main contribution in this paper.

3 Model Settings

3.1 Assumptions and Notations

We consider a three-player supply chain comprising a downstream manufacturer and two upstream suppliers. The upstream suppliers are heterogeneous, one is a small supplier with performance risk and capital constraint, and the other is a peer supplier who is capital sufficient and can always deliver products as ordered. The manufacturer can purchase from either the small or the peer supplier. For the remainder of this study, we use "he" to represent the small supplier, "she" to represent the manufacturer and "it" to represent the peer supplier.

We capture two important features of the small supplier in our model, which are common among suppliers seeking for financing. First, because of the small supplier's inherent unreliable, he will successfully deliver the products in a certain probability, i.e., performance risk. We use e to capture the small supplier's performance risk, where $e \in (0, 1)$. However, the small supplier can improve the delivery probability from 0 to e through unobservable effort which is costly. Specially, we assume the initial delivery probability to be 0. In order to increase the delivery probability from 0 to e, the small supplier has to afford costly effort of ke^2. We use k to capture the small supplier's operational efficiency: an efficient supplier has a smaller k. And the small supplier's operational level is common knowledge for all parties (the bank, downstream manufacturer and peer supplier).

Second, the small supplier's initial capital cannot cover his production cost. In order to release the financial stress, he will look for a loan. Specifically, we assume the effort cost is non-monetary and will not affect the small supplier's cash flow. What is more, the small supplier only has fixed assets. The financial constrained supplier can go to the commercial bank or downstream manufacturer or peer supplier for financing. The case that the small supplier finances from the commercial bank is POF, the case that the small supplier finances from the downstream manufacturer is BDF, and the case that the small supplier finances from the peer supplier is PF.

In some countries, the interest rate is controlled by external parties. We consider the interest rate of the loan is exogenous. We denote r_b as the bank's interest rate, r_m as the manufacturer's interest rate and r_p as the peer supplier's interest rate. The market demand faced by the manufacturer is assumed to be known. Specially, the demand is normalized to be 1. To cover the demand, the manufacturer procures from the upstream suppliers.

Before contracting stage, the manufacturer audits the small supplier's asset that is used as collateral. The value of the collateral asset is a, and is revealed to all parties. To avoid trivial cases, we assume that the small supplier's production cost c is no smaller than his collateral asset, i.e., $c > a$. In the contracting stage, the manufacturer decides the contract terms (including the small supplier's wholesale price) and the small supplier decides whether to accept it or not.

In the case when the small supplier delivers the products as ordered, the manufacturer pays the contract price w_1 and pays 0 otherwise. When the small supplier fails to deliver, the manufacturer purchases form the peer supplier at a wholesale price w_2, where $w_2 > w_1$. The peer supplier always delivers products as ordered.

3.2 Sequence of Events

- First, the peer supplier announces the wholesale price w_2;
- After learning w_2, the manufacturer audits the small supplier and decides the contract terms including the small supplier's wholesale price w_1;
- After accepting the order, the small supplier determines his effort e and goes to finance for production;
- The suppliers then deliver products to the manufacturer after production;
- Finally, the manufacturer sells products to the market.

4 Purchase Order Financing

In this scenario, the small supplier goes to the commercial bank for financing, i.e., *purchase order financing*. After accepting the purchase contract, the small supplier borrows c from the bank with the purchase order. With the loan, the small supplier begins production. The small supplier pays back the principle and interest, i.e., $(1 + r_b)c$, if he successfully delivers the order at the probability of e; otherwise, the small supplier pays back a (value of the collateral asset) at the probability of $1-e$. Further, we have the financing cost for the small supplier $e(1 + r_b)c + (1 - e)a - c$.

Thus, the small supplier's expected profit function is

$$\pi_s^{POF} = ew_1 - (e(1 + r_b)c + (1 - e)a) - ke^2 \tag{1}$$

where the first item is his revenue, the second item is the total cost of production and financing, the third item is the cost of effort.

The expected profit function for the peer supplier and the manufacturer can be formulized as below:

$$\pi_m^{POF} = p - ew_1 - (1-e)w_2 \tag{2}$$

$$\pi_p^{POF} = (1-e)(w_2 - c) \tag{3}$$

4.1 The Small Supplier's Effort Under POF

By backward induction and considering the first-order condition of (1), we obtain the small supplier's equilibrium effort level as

$$e(w_1) = \frac{a - c(1+r_b) + w_1}{2k} \tag{4}$$

Inserting (4) into (1), we can rewrite the small supplier's expected profit as

$$\pi_s^{POF} = \frac{a^2 + (c(1+r_b) - w_1)^2 - 2a(c(1+r_b) - w_1 + 2k)}{4k}$$

For the small supplier's participation constraint, i.e., $\pi_s^{POF} \geq 0$, we have

$$w_1 \geq c(1+r_b) + 2\sqrt{ak} - a \tag{5}$$

4.2 The Manufacturer's Decision Under POF

Inserting (4) into (2), we can rewrite the manufacturer's expected profit as

$$\pi_m^{POF} = p - w_2 + \frac{(a - c(1+r_b) + w_1)(w_2 - w_1)}{2k} \tag{6}$$

Then manufacturer's problem is to maximize her expected profit given by (6) under the constraint given by (5).

Thus, we have

$$\begin{cases} w_1(w_2) = \frac{w_2 + c(1+r_b) - a}{2}, \text{ if } w_2 > c(1+r_b) + 4\sqrt{ak} - a \\ w_1 = c(1+r_b) + 2\sqrt{ak} - a, \text{ otherwise} \end{cases} \tag{7}$$

4.3 The Peer Supplier's Decision Under POF

We insert the optimal wholesale price w_1 given by (7) into (4). Thus, the small supplier's optimal effort can be rewritten as

$$\begin{cases} e(w_2) = \frac{a-c(1+r_b)+w_2}{4k}, \text{ if } w_2 > c(1+r_b) + 4\sqrt{ak} - a \\ e = \sqrt{a/k}, \text{ otherwise} \end{cases} \tag{8}$$

Inserting the optimal effort given by (8) into the peer supplier's expect profit function given by (3), we have Proposition 1.

Proposition 1 *If* $0 < cr_b < a + 4k - 8\sqrt{ak}$*, we have the equilibrium under POF:* $e^{POF*} = \frac{a+4k-cr_b}{8k}$, $w_1^{POF*} = c + k + \frac{3(cr_b-a)}{4}$,$w_2^{POF*} = c + 2k + \frac{cr_b-a}{2}$
and the corresponding payoffs are

$$\pi_p^{POF*} = (4k + cr_b - a)^2/16k,$$

$$\pi_s^{POF*} = ((a - cr_b)^2 - 56ak - 8kcr_b + 16k^2)/64k,$$

$\pi_m^{POF*} = ((a - cr_b)^2 + 24k(a - cr_b) + 32k(p - c) - 48k^2)/32k$; *otherwise, we have* $e^{POF*} = \sqrt{a/k}$*,*
$w_1^{POF*} = c(1 + r_b) + 2\sqrt{ak} - a, w_2^{POF*} = c(1 + r_b) + 4\sqrt{ak} - a.$

$$\pi_s^{POF*} = 0,$$

$$\pi_p^{POF*} = (1 - \sqrt{a/k})(4\sqrt{ak} + cr_b - a), \pi_m^{POF*} = 3a - 4\sqrt{ak} + p - c(1 + r_b).$$

Proof of proposition 1 is given in the Appendix. In Proposition 1, we find that the optimal profits of the small supplier and the manufacturer decrease in the bank's interest rate r_b. Differently, the peer supplier's optimal profit increase in r_b. If the bank's interest rate is below a certain threshold, i.e., $0 < r_b < (a + 4k - 8\sqrt{ak})/c$, the small supplier's optimal effort level decreases and the peer supplier's optimal wholesale price increases with the bank's interest rate, which bring benefits to the peer supplier. Otherwise, r_b no longer affects the small supplier's optimal effort level. Comparing the two cases in Proposition 1, we can see that with POF, the bank's interest rate only affects the small supplier's optimal decision within a certain range $0 < r_b < (a + 4k - 8\sqrt{ak})/c$. Unlike the small supplier, the bank's interest rate always affects the optimal decision of the manufacturer and the peer supplier with POF.

5 Buyer Direct Financing

In this scenario, the downstream manufacturer supports the small supplier to help to mitigate his financial stress. The manufacturer may suffer from providing financing when the small supplier fails to deliver. However, the financing supporting can help to mitigate the small supplier's financial stress that reduces the manufacturer's purchase cost, which will incentive the manufacturer to be a lender. Thus, we have the two suppliers' expected profit as follows:

$$\pi_{\mathrm{s}}^{BDF} = ew_1 - (e(1 + r_m)c + (1 - e)a) - ke^2 \tag{9}$$

$$\pi_p^{BDF} = (1 - e)(w_2 - c) \tag{10}$$

Combining the features of buyer and lender, we formulate the manufacturer's expected profit as follows:

$$\pi_m^{BDF} = p - (ew_1 + (1 - e)w_2) + (e(1 + r_m)c + (1 - e)a - c) \tag{11}$$

where the first term is the sales revenue, the second term is the cost for procurement and the last term is the profit for providing financing.

5.1 The Small Supplier's Effort Under BDF

Considering the first-order condition of the small supplier's expected profit given by (9), we obtain

$$e(w_1) = a - c(1 + r_m) + w_1 \big/ 2k. \tag{12}$$

For the small supplier's participation constraint, i.e., $\pi_s^{BDF} \geq 0$, we have $w_1 \geq c(1 + r_m) + 2\sqrt{ak} - a$.

5.2 The Manufacturer's Decision Under BDF

Inserting (12) into (11), we can rewrite the manufacturer's expected profit as

$$\begin{aligned} \pi_m^{BDF} = &\frac{a - c(1 + r_m) + w_1}{2k}(w_2 - w_1 + c(1 + r_b) - a) \\ &+ p - w_2 + a - c \end{aligned} \tag{13}$$

Maximizing (13) under $w_1 \geq c(1+r_m)+2\sqrt{ak}-a$, we have the manufacturer's optimal decision as below

$$\begin{cases} w_1(w_2) = \frac{w_2}{2} + c(1+r_m) - a, \text{ if } w_2 > 4\sqrt{ak} \\ w_1 = c(1+r_m) + 2\sqrt{ak} - a, \text{ otherwise} \end{cases} \tag{14}$$

5.3 The Peer Supplier's Decision Under BDF

We insert (14) into (12).Thus, the small supplier's optimal effort can be rewritten as follows

$$\begin{cases} e(w_2) = \frac{w_2}{4k}, \text{ if } w_2 > 4\sqrt{ak} \\ e = \sqrt{a/k}, \text{ otherwise} \end{cases} \tag{15}$$

Inserting (15) into the expected profit function of the peer supplier given by (10), we draw the equilibrium of BDF in Proposition 2.

Proposition 2 *(i) For* $0 < c(1+r_m) < a + 2\sqrt{ak}$*, we obtain the equilibrium under BDF:*$e^{BDF*} = \sqrt{a/k}$*,* $w_1^{BDF*} = c(1+r_m) + 2\sqrt{ak} - a, w_2^{BDF*} = 4\sqrt{ak}.$

and the optimal payoffs $\pi_s^{BDF*} = 0,$

$\pi_p^{BDF*} = (1 - \sqrt{a/k})(4\sqrt{ak} - c),$

$\pi_m^{BDF*} = 3a - 4\sqrt{ak} + p - c.$

(ii) Otherwise, we obtain $e^{BDF*} = (c+4k)/8k,$

$$w_1^{BDF*} = c(1+r_m) + k + c/4 - a, w_2^{BDF*} = c/2 + 2k.$$

$$\pi_p^{BDF*} = (c-4k)^2/16k, \pi_s^{BDF*} = (c^2 + 8ck + 16k(k-4a))/64k,$$

$$\pi_m^{BDF*} = (32k(a+p) - 40ck + c^2 - 48k^2)/32k.$$

Proof of Proposition 2 is similar to that of Proposition 1 and is omitted for brevity. In Proposition 2, we can see that the manufacturer's interest rate will not influence both the optimal decisions of the small supplier and the peer supplier, i.e., r_m has no impact on the small supplier's effort e and the peer supplier's wholesale price w_2. Obviously, the small supplier has a higher optimal profit when $k + a + c/4 > c(1+r_m) \geq a + 2\sqrt{ak}$. In this case, we find the manufacturer also gains more.

6 Peer Financing

In this scenario, we study the efficiency of the supply chain coopetition. The small supplier goes to the peer supplier for financing. The peer supplier's expected profit is

$$\pi_p^{PF} = (e(1 + r_p)c + (1 - e)a - c) + (1 - e)(w_2 - c) \tag{16}$$

where the first item is the profit of financing the small supplier and the second item is the profit of fulfilling the manufacturer's order. The expected profit of the small supplier and manufacturer are as follows

$$\pi_s^{PF} = ew_1 - (e(1 + r_p)c + (1 - e)a) - ke^2 \tag{17}$$

$$\pi_m^{PF} = p - ew_1 - (1 - e)w_2 \tag{18}$$

6.1 The Small Supplier's Effort Under PF

Similarly, the small supplier's optimal effort is

$$e(w_1) = (a - c(1 + r_p) + w_1)\big/2k.$$

The participation constraint is

$$w_1 \geq c(1 + r_p) + 2\sqrt{ak} - a.$$

6.2 The Manufacturer's Decision Under PF

Inserting the small supplier's optimal effort into (18), we can rewrite the manufacturer's profit as

$$\pi_m^{PF} = (w_2 - w_1)\frac{a - c(1 + r_p) + w_1}{2k} + p - w_2 \tag{19}$$

Then the manufacturer's aim is to maximize (19) under $w_1 \geq c(1+r_p)+2\sqrt{ak}-a$. Thus, the manufacturer's optimal response is

$$\begin{cases} w_1(w_2) = \frac{w_2+c(1+r_p)-a}{2}, \text{ if } w_2 > c(1+r_p) + 4\sqrt{ak} - a \\ w_1 = c(1+r_p) + 2\sqrt{ak} - a, \text{ otherwise} \end{cases}$$

Then the small supplier's optimal effort can be rewrite as

$$\begin{cases} e(w_2) = \frac{a-c(1+r_p)+w_2}{4k}, \text{ if } w_2 > c(1+r_p) + 4\sqrt{ak} - a \\ e = \sqrt{a/k}, \text{ otherwise} \end{cases}$$

6.3 The Peer Supplier's Decision Under PF

Inserting the small supplier's optimal effort level above into (16), we conclude the equilibrium in Proposition 3.

Proposition 3 *If* $0 < c < 8\sqrt{ak} - 4k$,*we have the equilibrium and optimal payoffs under PF:*$e^{PF*} = \sqrt{a/k}$,$w_1^{PF*} = c(1 + r_p) + 2\sqrt{ak} - a$, $w_2^{PF*} = c(1 + r_p) + 4\sqrt{ak} - a$.

$$\pi_s^{PF*} = 0,$$

$$\pi_p^{PF*} = (1 - \sqrt{a/k})(4\sqrt{ak} - c) + cr_p,$$

$$\pi_m^{PF*} = 3a - 4\sqrt{ak} + p - c(1 + r_p).$$

Otherwise, we have $e^{PF*} = (c + 4k)/8k$,

$$w_1^{PF*} = c(1 + r_p) + k + \frac{c}{4} - a, w_2^{PF*} = c(1 + r_p) + 2k + \frac{c}{2} - a.$$

$$\pi_p^{PF*} = (c - 4k)^2/16k + cr_p,$$

$$\pi_s^{PF*} = (c^2 + 8ck + 16k(k - 4a))/64k,$$

$$\pi_m^{PF*} = (32k(a + p) - 40ck + c^2 - 48k^2)/32k - cr_p.$$

Proof of Proposition 3 is similar to that of Proposition 1 and is omitted for brevity. Unlike BDF scenario, in PF scenario the peer supplier's interest rate has impact on both suppliers' wholesale price as well as the optimal profits of the peer supplier and the manufacturer. Obviously, we can see that the wholesale price w_1^{PF*} and

w_2^{PF*} both increase in the peer supplier's interest rate r_p. As r_p increases, the peer supplier's optimal profit increases, while the manufacturer's optimal profit decreases. By comparing the peer supplier's profit in BDF and PF scenario, we find that the peer supplier benefits from supporting the small supplier, i.e., the peer supplier gains more because of coopetition.

Similar to BDF, the interest rate in PF has no impact on the small supplier's optimal effort. Comparing POF, BDF and PF, we can see that the small supplier's effort decision is only affected in POF scenario, i.e., the interest rate in internal financing schemes has no impact on the small supplier's effort. In other words, the small supplier's optimal decision is not affected by internal lenders' interest rate.

7 The Optimal Financing Decision

By comparing POF, BDF and PF scenarios, we address the small supplier's optimal financing scheme in Proposition 4.

Proposition 4 *(a) When* $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$, $0 < r_m < \frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k}$ *we have*:

(i) *for* $c \in (0, 8\sqrt{ak} - 4k)$,$\pi_s^{POF*} > \pi_s^{PF*} = \pi_s^{BDF*}$;
(ii) *for* $c \in (8\sqrt{ak} - 4k, \frac{a+2\sqrt{ak}}{1+r_m})$,$\pi_s^{PF*} > \pi_s^{P0F*} > \pi_s^{BDF*}$;
(iii) $\pi_s^{PF*} = \pi_s^{BDF*} > \pi_s^{POF*}$, *otherwise.*

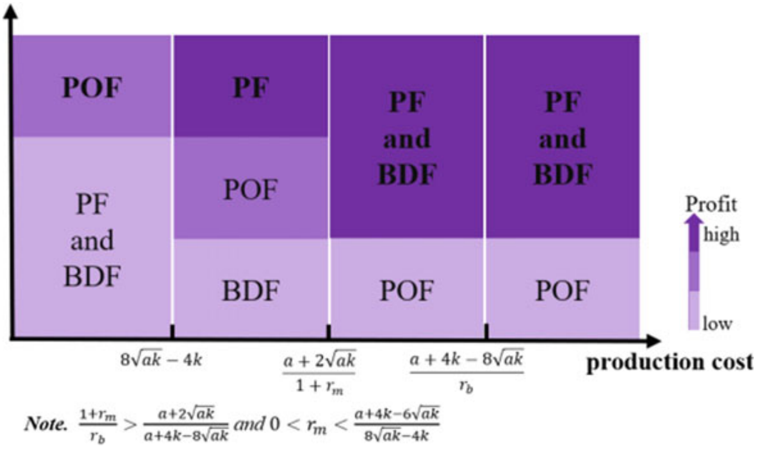
(b) When $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$, $\frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k} < r_m < 1$,*we have:*

(i) for $c \in (0, \frac{a+2\sqrt{ak}}{1+r_m})$,$\pi_s^{POF*} > \pi_s^{PF*} = \pi_s^{BDF*}$;
(ii) for $c \in (\frac{a+2\sqrt{ak}}{1+r_m}, 8\sqrt{ak} - 4k)$,$\pi_s^{BDF*} > \pi_s^{P0F*} > \pi_s^{PF*}$;
(iii) $\pi_s^{PF*} = \pi_s^{BDF*} > \pi_s^{POF*}$, otherwise.

(c) When $\frac{1+r_m}{r_b} < \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$,*we have:*

(i) or $c \in (0, 8\sqrt{ak} - 4k)$,$\pi_s^{POF*} > \pi_s^{PF*} = \pi_s^{BDF*}$;
(ii) for $c \in (8\sqrt{ak} - 4k, \frac{a+4k-8\sqrt{ak}}{r_b})$,$\pi_s^{PF*} > \pi_s^{P0F*} > \pi_s^{BDF*}$;
(iii) for $c \in (\frac{a+4k-8\sqrt{ak}}{r_b}, \frac{a+2\sqrt{ak}}{1+r_m})$,$\pi_s^{PF*} > \pi_s^{P0F*} = \pi_s^{BDF*}$;
(iv) $\pi_s^{PF*} = \pi_s^{BDF*} > \pi_s^{P0F*}$, otherwise.

Proof of Proposition 4 is given in the Appendix. When the interest rates of the bank and the manufacturer satisfy $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$ and $0 < r_m < \frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k}$, we obtain Fig. 1. Obviously, we can see that POF dominates when production cost is below $8\sqrt{ak} - 4k$. PF always dominates when production cost is higher than $8\sqrt{ak} - 4k$. BDF dominates when production cost is greater than $\frac{a+2\sqrt{ak}}{1+r_m}$. PF and BDF performs

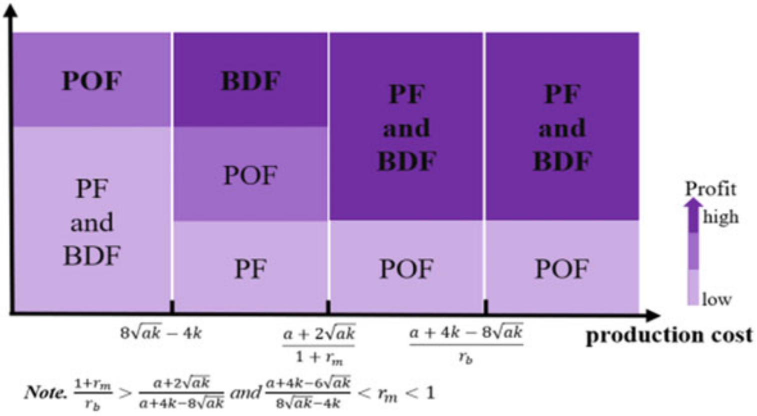


Note. $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$ *and* $0 < r_m < \frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k}$

Fig. 1 The small supplier's optimal financing scheme

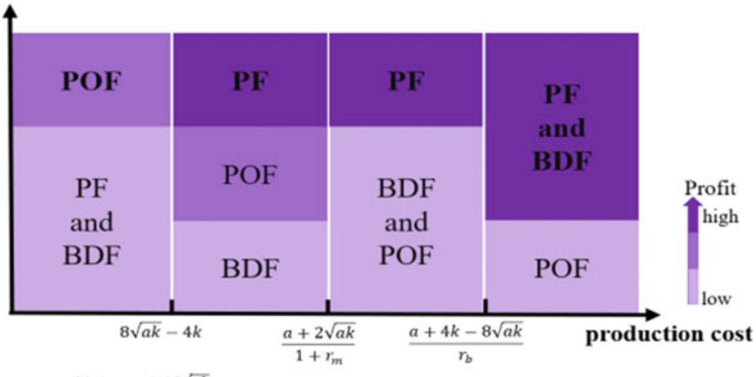
the same when the production cost breaks the threshold $\frac{a+2\sqrt{ak}}{1+r_m}$. Similarly, we have Figs. 2 and 3.

The figures given above summarize Proposition 4 and illustrate that when production cost is low the optimal financing scheme for the small supplier is the external financing, i.e., POF, otherwise the small supplier should choose internal financing, i.e., BDF or PF.



Note. $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$ *and* $\frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k} < r_m < 1$

Fig. 2 The small supplier's optimal financing scheme



Note. $\frac{1+r_m}{r_b} < \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$

Fig. 3 The small supplier’s optimal financing scheme

8 Conclusion

In this paper, we consider a capital-constrained supplier who can borrow money from the commercial bank or the downstream manufacturer or the peer supplier. We find that the performance-risk supplier’s effort level is only affected by the external lender’s interest rate (the bank’s interest rate). By comparing the equilibrium of three financing schemes (POF, BDF, PF), we find that the small supplier’s optimal financing decision is to go to the bank for financing if the production cost is low, otherwise to seek financing among supply chain members (buyer direct financing or peer financing). We also investigate the efficiency of coopetition in the supply chain. We find that the peer supplier’s financing support can also help to alleviate the small supplier’s financial stress. And, the peer supplier gains more because of the coopetition.

As the first attempt to investigate the performance of POF, BDF and PF, our paper has several limitations. First, the supply chain we considered is under symmetric information. Second, the supply chain members are all risk-neutral. Third, we use the interest rate as exogenous parameter. Fourth, the market demand in our model is assumed to be known and normalized to be 1.Future work can consider asymmetric information, risk-averse, demand uncertainty and endogenous the interest rate.

Acknowledgements This work is supported by national natural science foundation of china under grant 71771123.

Appendix: Proofs

Proof of Proposition 1

The firstand second-order condition of (1) is as follows: $\frac{\partial\pi_s^{BDF}}{\partial e} = a + w_1 - 2ek - c(1 + r_b)$ and $\frac{\partial^2\pi_s^{BDF}}{\partial e^2} = -2k < 0$. Thus, we have $e(w_1) = (a - c(1 + r_b) + w_1)\big/2k$.

For (6), its first and second-order condition is $\frac{\partial\pi_m^{POF}}{\partial w_1} = \frac{w_2-w_1-(a+w_1-c(1+r_b))}{2k}$, $\frac{\partial^2\pi_m^{POF}}{\partial w_1^2} = -\frac{1}{k} < 0$. Then, we obtain $w_1(w_2) = (w_2 + c(1 + r_b) - a)\big/2$. Considering the participation constraint $w_1 \geq c(1 + r_b) + 2\sqrt{ak} - a$, we have the manufacturer's optimal decision: $w_1(w_2) = (w_2 + c(1 + r_b) - a)\big/2$, if $\frac{w_2+c(1+r_b)-a}{2} \geq c(1 + r_b) + 2\sqrt{ak} - a$; $w_1 = c(1 + r_b) + 2\sqrt{ak} - a$, otherwise. Therefore, we have the small supplier's optimal decision in (8). As $0 < e < 1$, we have $a < k$. Inserting (8) into (3) we have.

$$\pi_p^{POF} = \{ {}^{\frac{a-c(1+r_b)+w_2}{4k}(w_2-c), w_2>c(1+r_b)+4\sqrt{ak}-a}_{(1-\sqrt{a/k})(w_2-c),\text{ otherwise}} .$$

If $w_2 > c(1 + r_b) + 4\sqrt{ak} - a$, we have $\frac{\partial\pi_p^{POF}}{\partial w_2} = \frac{4k-a+c(1+r_b)-2w_2+c}{4k}, \frac{\partial^2\pi_p^{POF}}{\partial w_2^2} = -\frac{1}{2k} < 0$, Thus $w_2 = c + 2k + \frac{cr_b-a}{2}$. Solving the inequation $w_2 = c + 2k + \frac{cr_b-a}{2} > c(1 + r_b) + 4\sqrt{ak} - a$, we have $cr_b < a + 4k - 8\sqrt{ak}$. Then, if $0 < cr_b < a + 4k - 8\sqrt{ak}$, we have the equilibrium: $w_1^{POF*} = c + k + \frac{3(cr_b-a)}{4}, e^{POF*} = \frac{a+4k-cr_b}{8k}, w_2^{POF*} = c + 2k + \frac{cr_b-a}{2}$; Otherwise, $\partial\pi_p^{POF} = 1 - \sqrt{\frac{a}{k}} > 0$, we have π_p^{POF} increases as w_2. From $w_2 \leq c(1 + r_b) + 4\sqrt{ak} - a$ we obtain $w_2 = c(1 + r_b) + 4\sqrt{ak} - a$, and further have the equilibrium:$w_2^{POF*} = c(1 + r_b) + 4\sqrt{ak} - a, w_1^{POF*} = c(1 + r_b) + 2\sqrt{ak} - a, e^{POF*} = \sqrt{a\big/k}$.

Proof of Proposition 4

Suppose $\frac{a+4k-8\sqrt{ak}}{r_b} > \frac{a+2\sqrt{ak}}{1+r_m} > 8\sqrt{ak} - 4k$, we have $\frac{1+r_m}{r_b} > \frac{a+2\sqrt{ak}}{a+4k-8\sqrt{ak}}$ and $r_m < \frac{a+4k-6\sqrt{ak}}{8\sqrt{ak}-4k}$.

(i) When $c \in (0, \frac{a+2\sqrt{ak}}{1+r_m})$, we have $\pi_s^{PF*} = \pi_s^{BDF*} = 0$, $\pi_s^{POF*} = ((a - cr_b)^2 - 56ak - 8kcr_b + 16k^2)/64k$. Thus, we have $\pi_s^{POF*} > \pi_s^{PF*} = \pi_s^{BDF*}$;

(ii) when $c \in (8\sqrt{ak} - 4k, (a + 2\sqrt{ak})/(1 + r_m))$, we have $\pi_s^{PF*} = [c^2 + 8ck + 16k(k - 4a)]/(64k)$. Further, we get $\pi_s^{PF*} - \pi_s^{POF*} = c^2 + 8ck(1 + r_b) - 8ak - (a - cr_b)^2$. From $c > a$, it follows that $\pi_s^{PF*} - \pi_s^{POF*} > 0$. Then we have $\pi_s^{PF*} > \pi_s^{P0F*} > \pi_s^{BDF*}$;

(iii) otherwise, we have $\pi_s^{BDF*} = [c^2 + 8ck + 16k(k - 4a)]\,/(64k)$, then we have $\pi_s^{PF*} = \pi_s^{BDF*} > \pi_s^{POF8}$.

For part (b) and (c), we could prove it similarly.

References

1. Report on Employer Firms Based on the 2018 Small Business Credit Survey, Fed Small Business (2019). https://www.fedsmallbusiness.org/survey/2019/report-on-employer-firms
2. C.H. Lee, B.D. Rhee, Trade credit for supply chain coordination. Eur. J. Oper. Res. **214**(1), 136–146 (2011)
3. P. Kouvelis, W. Zhao, The newsvendor problem and price-only contract when bankruptcy costs exist. Prod. Oper. Manage. **20**(6), 921–936 (2011)
4. P. Kouvelis, W. Zhao, Financing the newsvendor: supplier vs. bank, and the structure of optimal trade credit contracts. Oper. Res. **60**(3), 566–580 (2012)
5. B. Jing, X. Chen, G. Cai, Equilibrium financing in a distribution channel withcapital constraint. Prod. Oper. Manage. **21**(6), 1090–1101 (2012)
6. S.A. Yang, J.R. Birge, Trade credit, risk sharing, and inventory financing portfolios. Manage. Sci. **64**(8), 3667–3689 (2018)
7. T.I. Tunca, W. Zhu, Buyer intermediation in supplier finance. Manage. Sci. **64**(12), 1–20 (2017)
8. M. Reindorp, F. Tanrisever, A. Lange, Purchase order financing: credit, commitment, and supply chain consequences. Oper. Res. **66**(5), 1287–1303 (2015)
9. C.S. Tang, S.A. Yang, J. Wu, Sourcing from suppliers with financial constraints and performance risk. Manuf. Serv. Oper. Manage. **20**(1), 70–84 (2018)
10. Y. Alan, V. Gaur, Operational investment and capital structure under asset-based lending. Manuf. Serv. Oper. Manage. **20**(4), 637–654 (2018)
11. M.N. Katehakis, B. Melamed, J. Shi, Cash-flow based dynamic inventory management. Prod. Oper. Manage. **25**(9), 1558–1575 (2016)
12. P. Kouvelis, W. Zhao, Who should finance the supply Chain? Impact of credit ratings on supply Chain decisions. Manuf. Serv. Oper. Manage. **20**(1), 19–35 (2018)

Evaluating Transportation Routes Between China and Vietnam Based on Delphi–CFPR

Thi Phuong Thao Doan, Lixin Shen, Xiaowen Shi, Zaili Yang, Chuang Li, and Ke Jing

Abstract China-Vietnam trade significantly grows in the recent decades, and their trade value in 2019 arrives at approx. 117 billion USD. The fast growth trade stimulates the development of the transport infrastructure and new routes between two countries. Currently the China-Vietnam trade is implemented by road, rail, water, air and intermodal transportation. To respond to the strong demand of trade and taking into account the significant role of containerization in international trade, selecting an efficient intermodal route becomes necessary. This study aims to evaluate the competitive routes for 40 ft container transportation between China and Vietnam through an empirical study investigating the routes between Chongqing, China and Haiphong, Vietnam. A hybrid Delphi and CFPR method is applied with the support of quantitative and qualitative data to obtain the goal from logistics service provider and shipper's perspectives. The results reveal that among the eight main factors, transportation cost is the most important influencing route choice for 40 ft containers. Moreover, the eight multimodal routes from Chongqing to Haiphong are evaluated to have the best route identified as aviation. This research will benefit policymakers and state institutions by the provision of the available routes and their advantages

T. P. T. Doan
School of Transportation Engineering, Dalian Maritime University, Dalian, China
e-mail: doanthaovmu2608@gmail.com

L. Shen (✉) · X. Shi · C. Li · K. Jing
School of Maritime Economics and Management, Dalian Maritime University, Dalian, China
e-mail: shenlixin@dlmu.edu.cn

X. Shi
e-mail: shixiaowen@dlmu.edu.cn

C. Li
e-mail: braveli@dlmu.edu.cn

K. Jing
e-mail: jingke@dlmu.edu.cn

Z. Yang
Liverpool Logistics Offshore and Marine Research Institute, Liverpool John Moores Uniersity, Liveprool, UK
e-mail: z.yang@ljmu.ac.uk

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_34

against each key criterion. Furthermore, this research provides useful insights for logistics companies and shippers in transporting cargoes from China and Vietnam.

Keywords Belt and road initiative · Sino-vietnamese trade · Intermodal · CFPR · Delphi

1 Introduction

China and Vietnam are connected with full of transportation modes such as waterways, railway, road and aviation. In recent years, the demand for goods originating from China of Vietnamese people has increased rapidly, accounting for 35% of the national market. Therefore, transportation from China to Vietnam is entering the development. Roadway is one of the popular forms of transporting goods from China to Vietnam. The advantage of this form is flexible and convenient.

Figure 1 shows that it takes 3 days to move cargoes by trucks from Chongqing, China to Hanoi, Vietnam. This route is passing through Kunming, to cross the Hekou border. This is the key way to transport goods between the Northwest of China and Vietnam. The second route is from Shanghai, China to Hanoi, Vietnam, taking 4 days in total by trucks first from Shanghai Shenzhen and then to Hanoi through Nanning, Guangxi, China and cross Pingxiang Border.

Secondly, mentioning about rail transportation, according to the Vietnam Custom, the total volume of import products by railway from China to Vietnam was 453.5 million tons. The total volume of export products by railway from Vietnam to China was 387.6 million tons in 2017 (Tables 1, 2 and 3).

Sea transportation with large cargo will save costs. The cost of shipping is lower than that of other forms. The container is used by many shipping companies from

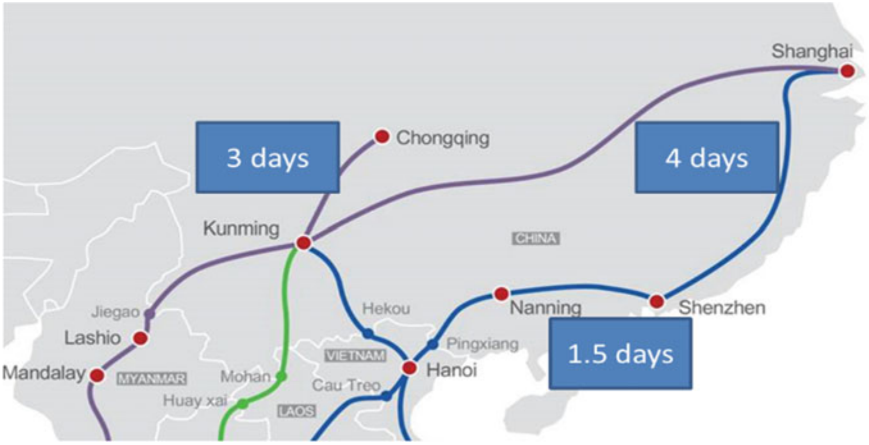


Fig. 1 Main transportation roads between China and Vietnam

Table 1 Main railway routes between China and Vietnam

Departure	Destination	Distance (km)	Time
Kunming	Haiphong	855	2 days
Beijing	Hanoi	2595	37 h 21 min
Nanning	Hanoi	396	11 h 25 min
Nanchang	Hanoi	1700	5 days
Chongqing	Hanoi		99 h

Table 2 Main sea transportation routes between China and Vietnam

Port	Time					
	Shanghai	Guangzhou	Ningbo	Qingdao	Tianjin	Xiamen
Ho Chi Minh	6	3	7	8	10	7
Hai Phong	4	2	8	6	8	4

Table 3 Linguistic terms for importance weight of the criteria

Relative importance	Linguistic terms
1	Equally important (EI)
2	Weakly more important (WI)
3	Strongly more important (SI)
4	Very strongly more important (VI)
5	Absolutely more important (AI)

China to Vietnam so it will avoid breakage and dirt. Commodities such as liquids are also not restricted.

According to the Civil Aviation Administration of China in 2016, China owns the busiest airports in the world such as Shanghai Pudong International Airport (handled 3,440,279.7 metric tons), Beijing Capital International Airport (handled 1,943,159.7 metric tons), Chongqing Jiangbei International Airport (handled 361,091.0 metric tons). There are several main airports in Vietnam such as Tan Son Nhat International Airport in the South (handled 430,627 metric tons of cargo in 2015), Noi Bai International Airport in the North (handled a total of 566, 000 metric tons of cargo in 2017), Da Nang International Airport in the Central of Vietnam.

Furthermore, according to [1] Vietnam Custom report, China is the biggest trade partner of Vietnam in 2019 with the turnover of 116.866 billion USD. As the fastest growing country in the world, China has been tried to improve its intermodal transport system for supporting efficient international trading based on Belt Road Initiative (BRI). In the past decades, China has been famous for dominating global production relating low manufacturing costs and wages. However, these foremost advantages of investors have increasingly diminished due to the mature of China's economy worldwide. As a result, Chinese manufacturers are expanding their business beyond

China, called the "China Plus one Strategy" to other Asian countries such as Vietnam, Myanmar, Thailand. Because developing countries are increasingly implementing [2] GVCs, China tends to export more intermediate goods to help their final goods exports to the global market.

Hence, with the strong demand of trading between two countries, the demand from BRI in decreasing transport time from China to ASEAN countries, it is necessary to evaluate route selection between China and Vietnam.

2 Literature Review

In the process of supply chain management, one of the most important decisions is transportation route selection to improve logistics efficiency and effectiveness [3]. Indeed, there are three major trends to apply different methods for the process of selecting transportation modes among transport researchers including of mathematical algorithms, cost–benefit analysis, and multi-criteria decision-making methods.

An optimal route qualifies to minimize costs and time to consider with some constraints by applying mathematical algorithms such as mixed integer programming and dynamic programming. Ayar and Yaman [4] studied a problem of multimodal routing using ground and maritime transport at a minimum cost and reserve by two mixed integer programming.

Moreover, a number of studies focused on specific route or industry to clarify the problem. Cho et al. [5] calculated both time and transport costs for actual multimodal transport routes from Busan, Korea to Rotterdam, Netherlands by applying Pareto optimal method. Xie et al. [6] specifically examined the problem in the transport of hazardous materials by multimodal transportation multiplying mixed integer programming. Bookbinder and Fox [7] have adopted the shortest path algorithm to determine a route that qualifies minimum time and cost from Canada to Mexico.

Other studies considered additional costs and other factors in shipping options such as storage costs and shipping costs by using mathematical formulas such as distribution costs, shipping time [8]. Samimi et al. [9] introduced two binary mode selection models to test transport modes between trucks and railroads to transport goods in the United States. Specific batch variables such as distance, weight and mode Specific variables such as transit time and cost are the influencing factors which affect truck and rail competition.

In addition, several studies have applied cost–benefit analysis approach as the classical economic model [10] to take into account the savings in transportation time and cost to test the experimental studies.

The time–cost–distance method will be simple by using graphical illustration of two variables calling cost and time to determine inefficiencies and bottlenecks of each route. Regmi and Hanaoka [11] examined international multimodal transport corridors between Northeast and Central Asia. The simple time–cost distance method was used to evaluate multimodal transport corridors by means of navigation, road and rail.

The study took a conclusion that infrastructure, cross-border processes, interactions of border transport modes, unavailability of wagons and the frequency of cargo ships were physical bottlenecks during testing. Meanwhile, trucks and wagons, excessive security checks of cargoes, the differences in border cross—border processes and opening working time of the control border offices were non—physical constraints.

Another method of solving routing problems is adherence to the MCDM process, which proposes that not only costs and time but also other factors such as reliability, security and safety of transport routes is significantly considered. Norojono and Young [12] applied the stated priority model for the selection of freight modes. The study identified hierarchical attributes including 4 main elements and 9 subfactors. The model has demonstrated that safety, reliability and responsiveness are the main determinants of the choice between rail and road in Java, Indonesia from the perspective of shipper. Moon et al. [13] analyzed the competitiveness of six intermodal routes from Korea to Europe including railway and seaway by using technology in order of priority by similarity to ideal solution (TOPSIS) and triangular fuzzy numbers. The priority of the alternatives has been decided based on both quantitative factors and qualitative factors. Wang and Yeo [14] applied Fuzzy Delphi to determine the optimal shipping route from South Korea to Central Asia for exporting secondhand cars. Combining experts' opinion and collected real data, three options were evaluated based on factors including total cost, total time, reliability and transport capacity. The result of this research showed that total cost is the most prioritized factor. Recently, Wang and Yeo [15] applied a hybrid method of Fuzzy Delphi Electre I to evaluate intermodal routes from Korea to Central Asia. The results improved that the cost factor was the most important one of the five major factors and among the sub-factors, the cost of transport and cooperation between state institutions were preferred.

From previous studies, it has been shown that multi-criteria decision-making methods can be applied in many industries. However, only a few studies have applied the combination of CFPR and Delphi in transport route analysis. Hence, the context of China–Vietnam transport routes remains unexplored. Therefore, the proposed CFPR and Delphi are implemented in this thesis to focus on analyzing multilateral routes for bilateral trade between China and Vietnam.

3 Methodology

In this research, the CFPR method was applied to solve the MCDM problem of transport route selection for a 40ft container. The Delphi method was employed to identify a hierarchy of criteria. Although a small sample was used, the Delphi method's results were objective and reasonable.

The CFPR method helped in evaluating the process of transport route selection by ranking the alternatives through computational simplicity and consistency. The CFPR method has also been used to solve decision-making problems in the literature, as illustrated in Fig. 2.

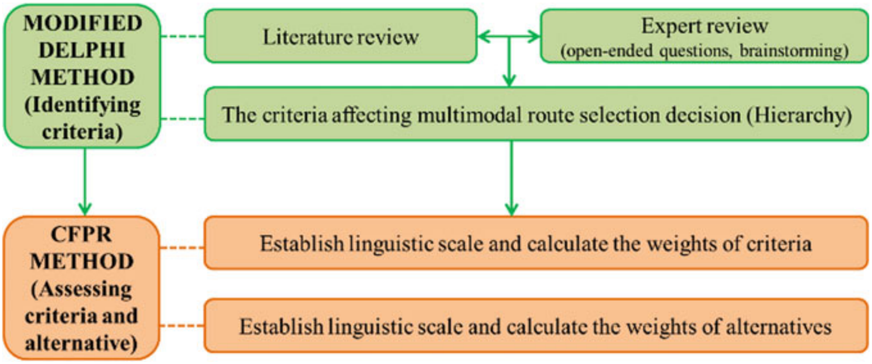

MODIFIED DELPHI METHOD (Identifying criteria)
Literature review
Expert review (open-ended questions, brainstorming)
The criteria affecting multimodal route selection decision (Hierarchy)
CFPR METHOD (Assessing criteria and alternative)
Establish linguistic scale and calculate the weights of criteria
Establish linguistic scale and calculate the weights of alternatives

Fig. 2 Schematic diagram of research

3.1 The Delphi Method

The Delphi method was firstly developed by RAND Corporation in the 1950s to predict the effectiveness of technology on war. The goal of the method is to achieve the most reliable outcome by answering the group for sequential questions to solve a complex problem [16]. Therefore, this method is applied to many areas as a technique to identify and prioritize management decision making issues [17–22].

The size of the Delphi group is determined by the expertise and dynamics of the expert council instead of statistical power to reach consensus among all participants [23]. Delbecq et al. [24] recommended that a group of five to nine experts are sufficient to achieve a reasonable assessment, and therefore the study used a decision-making group of seven experts.

3.2 The CFPR Method

The research applied the CFPR method developed by Herra-Viedma et al. [25] to evaluate the multilateral routes selection. The CFPR method establishes a priority decision matrix in pairwise comparison determined by the additive transition property [25]. The advantage of CFPR is that (1) significantly shortening the length of the questionnaire by reducing the number of questions to compare (n-1) for a group of n criteria leading to an increased possibility of receiving reliable responses; (2) avoiding inconsistency and reevaluation responses that not only saves time but is also more effective.

The study applied important definitions and recommendations proposed by Chen and Chao [26] with some modifications as a methodological improvement, (1) a relative importance scale reduced from 9 to 5 to allow simpler judgment, from seventeen magnitude importance of ripening by pair comparison; (2) evaluation criteria include quantitative and qualitative factors. In terms of quantitative factors, real data is collected and qualitative factors are collected by questionnaires.

4 Empircal Analysis

The main purpose of the research is to evaluate and analyze the possible alternative transport routes for importing a 40 ft container from China to Vietnam via multimodal transportation modes. To make sure a fair comparison, Chongqing and Haiphong cities are selected for an origin and a destination respectively for several reasons. Indeed, Chongqing, as an origin, and Haiphong, as a destination, were determined because the two cities are the main logistic hubs in China and Vietnam, respectively.

Keeping inside geographically strategic locations and being economic hubs, Chongqing and Haiphong were well connected by major transport modes such as marine, inland-waterway airway, railway and roadway. The eight common alternative routes are examined in the study, as shown in Fig. 3:

Alternative 1: Chongqing–Hanoi–Haiphong (Airway)—R1
Alternative 2: Chongqing–Shanghai–Haiphong (Road + Ocean)—R2
Alternative 3: Chongqing–Shanghai–Haiphong (Rail + Ocean)—R3
Alternative 4: Chongqing–Shanghai–Haiphong (Inland waterway + Ocean)—R4
Alternative 5: Chongqing–Shenzhen–Haiphong (Road + Ocean)—R5

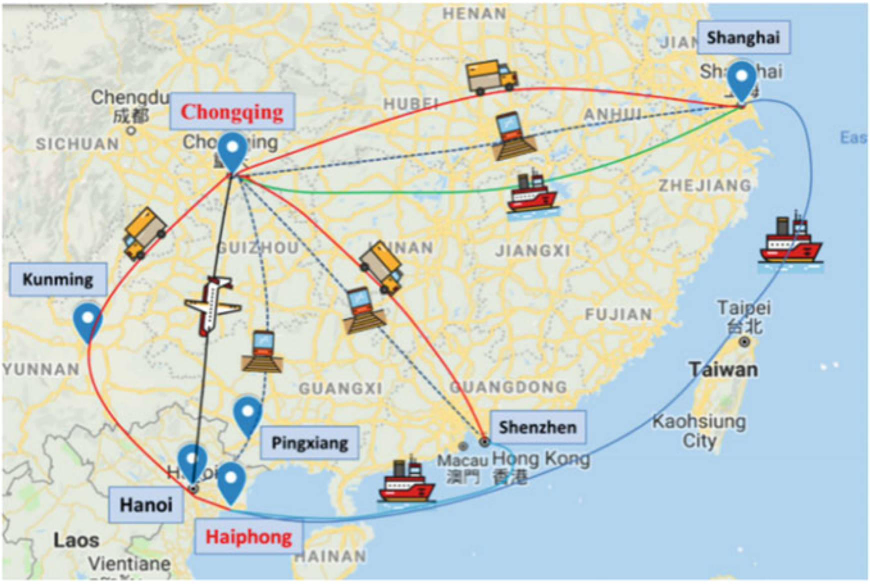


Fig. 3 The alternative routes for transporting a 40 ft container from Chongqing to Haiphong

Alternative 6: Chongqing–Shenzhen–Haiphong (Rail + Ocean)—R6
Alternative 7: Chongqing–Haiphong (Road)—R7
Alternative 8: Chongqing–Hanoi–Haiphong (Rail + Road)—R8.

4.1 The Survey Design

To reach the research's goals, two steps of methodology were designed (Fig. 4).

Initially, Delphi method was applied to obtain a hierarchy of main factors and sub-factors affecting to multimodal route selection. In the first step, factors were determined through literature reviews and open-ended questions with related-field experts. Subsequently, the research applied CFPR method to evaluate and rank the factors and alternatives. The main factors and sub-factors system in expert's knowledge was defined through likert scale. Then, the alternatives were compared to determine the optimal option.

1. *Identifying Factors and Alternatives*

In the first step of the research, literature relating to route selections and transport modes was circulated among related-field experts for five round to gain the last result. A group of seven experts were invited to answer the questionnaire. To be more detail, all of experts are from huge enterprises (including of C and D enterprise, Yangming, Evergreen, Wanhai lines, ONE line, COSCO, APL). The working experiences of experts are from 5 to 20 years in related field. In addition, the experts' positions are

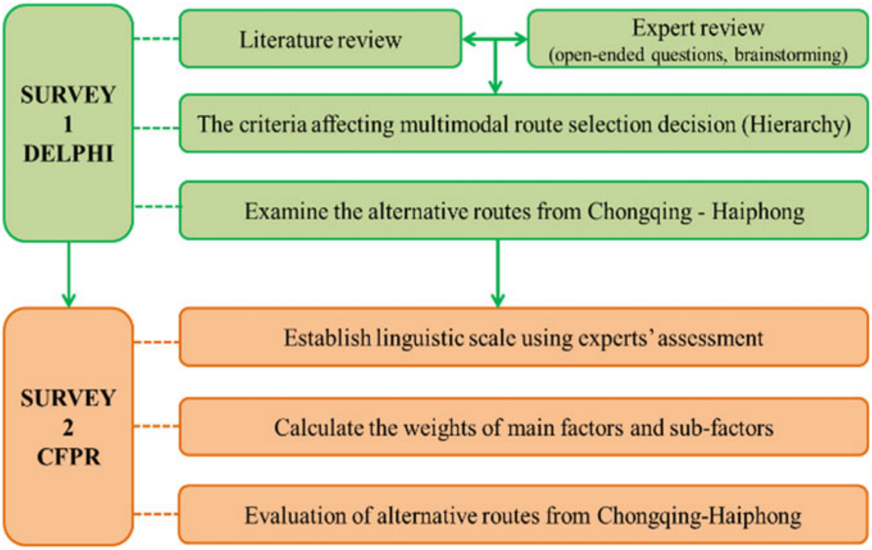

SURVEY
1
DELPHI
Literature review
Expert review
(open-ended questions, brainstorming)
The criteria affecting multimodal route selection decision (Hierarchy)
Examine the alternative routes from Chongqing - Haiphong
SURVEY
2
CFPR
Establish linguistic scale using experts' assessment
Calculate the weights of main factors and sub-factors
Evaluation of alternative routes from Chongqing-Haiphong

Fig. 4 The two-step methodology procedure

abundant such as purchasing staffs, operation staffs, operation managers, customer service staffs. The surveys were collected over one month from May 2019 to June 2019 through email, cellphones and face-to-face interviews.

The expert would answer about the factors determined from reviewing related literatures. To be more specific, those experts were asked to identify whether factors were necessary and whether factors overlapped through answering open-ended questions. The questionnaire was collected in sequential rounds to revise the experts' opinions. After this step, a system of factors was defined including of eight main factors and twenty-five sub-factors, alongside eight alternatives. (Fig. 5).

2. *Weighting Factors and Alternatives*

In this step, the experts evaluated the main factors and sub—factors get on the first step. To be more detail, the seven experts participating in the first round survey were continue to join in this part, and eight additional experts from electronics manufacturers such as C and D enterprise, Changhong enterprise, Yonghan enterprise and so on were also invited to answer the questionnaire in the second survey. Totally, there were 15 respondents with the experiences of experts is from 3 to 20 years in related fields. The number of respondents was adequate for an in-depth interview with experienced expert. The second survey was conducted over 50 days, from July 2019 to September 2019.

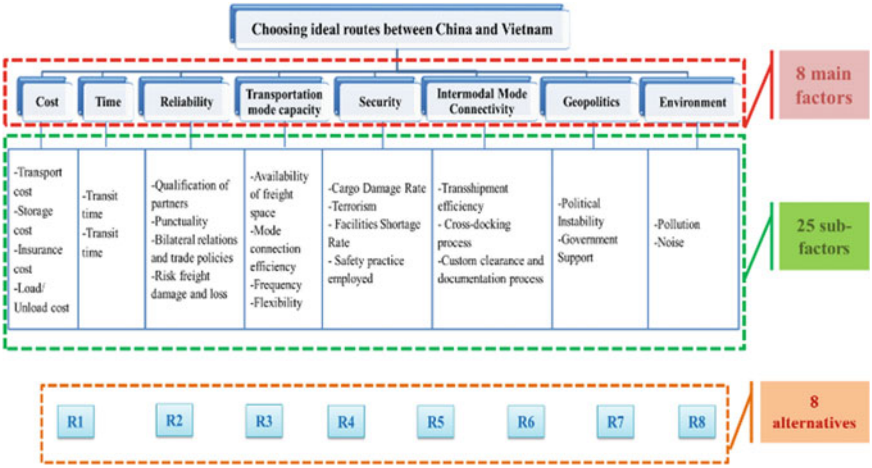

Choosing ideal routes between China and Vietnam
Cost
Time
Reliability
Transportation mode capacity
Security
Intermodal Mode Connectivity
Geopolitics
Environment
8 main factors
-Transport cost
-Storage cost
-Insurance cost
-Load/ Unload cost
-Transit time
-Transit time
-Qualification of partners
-Punctuality
-Bilateral relations and trade policies
-Risk freight damage and loss
-Availability of freight space
-Mode connection efficiency
-Frequency
-Flexibility
-Cargo Damage Rate
-Terrorism
- Facilities Shortage Rate
- Safety practice employed
-Transshipment efficiency
- Cross-docking process
-Custom clearance and documentation process
-Political Instability
-Government Support
-Pollution
-Noise
25 sub-factors
R1
R2
R3
R4
R5
R6
R7
R8
8 alternatives

Fig. 5 The structure of multi-criteria for route selection from China to Vietnam

4.2 Evaluation of the Criteria and Alternative Routes from Chongqing to Haiphong

1. *Weighting Calculations and Evaluating the Criteria*

Pair comparisons for the eight principal factors and 25 sub-factors employed in the CFPR methodology were performed (Fig. 6).

The results illustrated that, among the eight main factors, transportation cost is prioritized when considering the transport routes for a 40 ft container, followed by reliability, transportation time and security. The least priority factor is geopolitics. Meanwhile, of the 25 sub-factors, transport time is the most important factor. Therefore, the airway route is preferred over the seven other alternatives to ship from Chongqing to Haiphong. In contrast, the least important sub-factor is flexibility.

To be more detail, with respect to the main factors, the results for the eight principal factors indicated that transportation cost ranked first with 17.3%. According to the logistics companies and shippers, transportation cost is the most important criteria to decide the transport route. This finding is consistent with the required features for most products that prioritize minimizing cost. However, reliability is a crucial determinant when selecting transportation modes. This result is suitable with the high-value cargoes which mainly require safety, in which qualification of partners, punctuality, bilateral relations and trade policies, risk freight damage and loss are sub factors. Furthermore, to ensure seamless transport, transportation time and security is

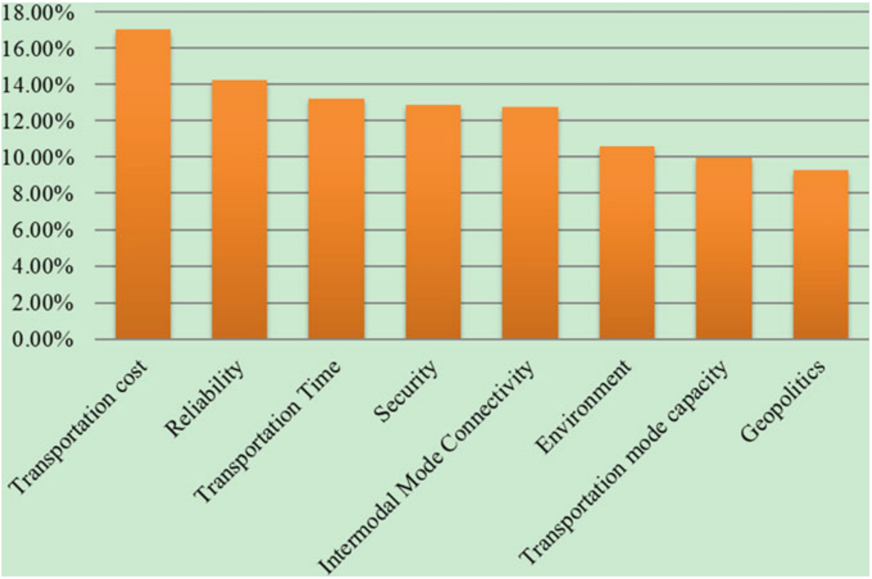


Fig. 6 The weight of main criterias

also a vital factor that must be considered, including transport time, transit time, cargo damage rate, terrorism, facilities shortage rate, safety practice employed (Fig. 7).

Of the sub-factors, transportation time is the most important. For transporting high-value cargo, it is essential to be fast because the price of high-tech products change day by day. Among the pollution, transport cost, storage cost, transshipment efficiency gains the most attention. Indeed, that would consistently explain because, on average, transportation cost accounts for 20% of manufacturing companies' total production costs and more than 50% of the logistics costs for transportation. Additionally, political instability and risk freight damage and loss is ranked sixth and seventh, as international transport depends on more than one country's administration. Thus, the involved government departments, trade policies, and political relationships directly and indirectly influence route development. In addition, among the manufacturing inputs, transit time of delivery is pivotal because as analyzed above the importance of time in transporting high-tech products. Similarly, due to the characteristics of transported cargo, cross-docking process is important criteria. Frequency is also vital factors for transport selection. The more frequent are the transport services provided, the lower is the inventory level the shippers require, resulting in lower total costs. The higher is flexibility the service provides, the lower are the losses suffered if there are changes or delays due to uncertainty in the supply chain.

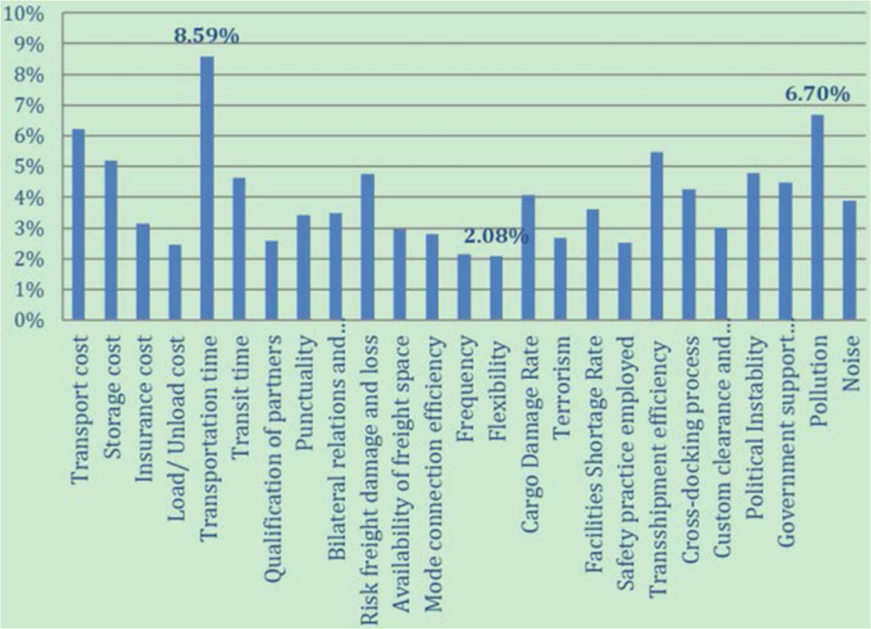


Fig. 7 The weight of sub criterias

2. *Weighting Calculation and Evaluating the Alternatives with Respect to the Main Factors*

In this research, the eight alternatives for importing a 40 ft container from Chongqing to Haiphong were evaluated based on eight main factors. The eight main factors divided into two main groups: six qualitative factors and two quantitative factors. The two quantitative factors (transportation cost and transportation time) were collected to calculate the objective factors (Table 5), While the subjective factors were collected through the experts' opinions from the second questionnaire.

With the objective factors, collected data with different units would be transformed into dimensionless indices to be suitable compatibility with the linguistics variables of the subjective factors. For quantitative factors, the higher the value is, the lower the competitiveness (cost, time and distance). The alternative with the highest benefit (or the lowest cost) would receive the highest score (Table 6).

The overall ranks considering both quantitative and qualitative methods are Route1 > Route3 > Route6 > Route4 > Route2 > Route7 > Route5 > Route8 as in Fig. 8.

Table 4 illustrates the evaluation of the eight alternatives, and it indicates the airway route was preferred to transport from Chongqing to Haiphong, followed by the rail and ocean route through Shanghai port and then the rail and ocean route through Shenzhen port. The least preferred route belonged to the combine of rail and road mode with the weight of only 0.112. To sum up, according to the shippers, a service by airway is preferred. Although airway is the most expensive option among the eight alternatives, the outstanding advantage is the high reliability, speed, and security, especially with the high-value products.

Taking a deep look to the quantitative factors table showed that, the airway alternative have advantage in time. It took only about one day for transporting. Meanwhile, the forth alternative (Chongqing–Shanghai–Haiphong by Inland waterway

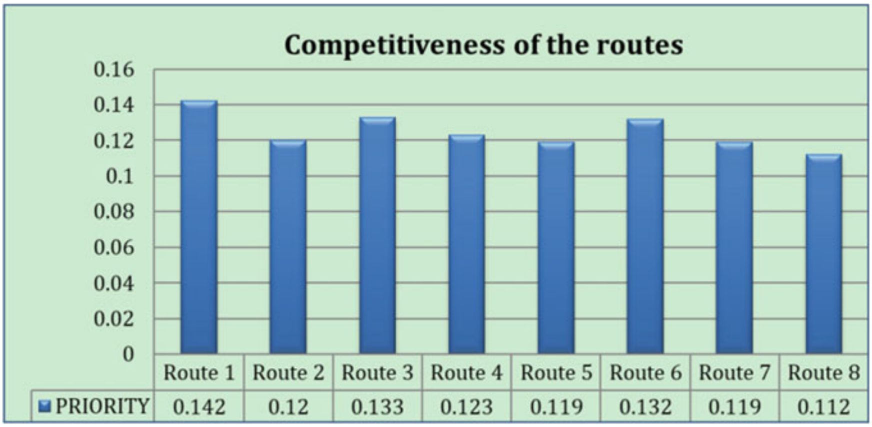


Fig. 8 The score and rank of the alternatives with respect to each main-factor

Table 4 The weight and rank of main and sub-criteria

Main factors	Local importance	Ranking	Sub factors	Local importance	Global importance	Ranking
Transportation cost	0.170	1	Transport cost	0.37	0.062	3
			Storage cost	0.31	0.052	5
			Insurance cost	0.18	0.031	16
			Load/ Unload cost	0.14	0.025	23
Transportation time	0.132	3	Transport time	0.65	0.086	1
			Transit time	0.35	0.046	8
Reliability	0.143	2	Qualification of partners	0.18	0.026	21
			Punctuality	0.24	0.034	15
			Bilateral relations and trade policies	0.24	0.035	14
			Risk freight damage and loss	0.33	0.048	7
Transportation mode capacity	0.100	7	Availability of freight space	0.30	0.030	18
			Mode connection efficiency	0.28	0.028	19
			Frequency	0.21	0.021	24
			Flexibility	0.21	0.021	25
Security	0.129	4	Cargo damage rate	0.32	0.041	11
			Terrorism	0.21	0.027	20
			Facilities shortage rate	0.28	0.036	13
			Safety practice employed	0.20	0.025	22
Intermodal mode connectivity	0.128	5	Transshipment efficiency	0.43	0.055	4
			Cross-docking process	0.33	0.043	10
			Custom clearance and documentation process	0.24	0.030	17
Geopolitics	0.093	8	Political instablity	0.52	0.048	6

(continued)

Table 4 (continued)

Main factors	Local importance	Ranking	Sub factors	Local importance	Global importance	Ranking
			Government support corruption	0.48	0.045	9
Environment	0.106	6	Pollution	0.63	0.067	2
			Noise	0.37	0.039	12

Table 5 Data of the objective factors

	Transportation routes	Cost (USD/40HQ)	Time (Day)
1	Chongqing–Hanoi–Haiphong (Airway)—R1	140,000	0.9
2	Chongqing–Shanghai–Haiphong (Road + Ocean)—R2	2373	9
3	Chongqing–Shanghai–Haiphong (Rail + Ocean)— R3	1897.6	9.5
4	Chongqing–Shanghai–Haiphong (Inland waterway + Ocean)—R4	1089.6	20
5	Chongqing–Shenzhen–Haiphong (Road + Ocean)— R5	2089.5	6
6	Chongqing–Shenzhen–Haiphong (Rail + Ocean)— R6	1862	6.5
7	Chongqing–Haiphong (Road)—R7	2950	4
8	Chongqing–Hanoi–Haiphong (Rail + Road)—R8	2170	5

Table 6 Competitiveness of the routes considering both quantitative and qualitative methods

Main factor	R1	R2	R3	R4	R5	R6	R7	R8
Reliability	0.022	0.017	0.019	0.016	0.016	0.022	0.019	0.014
Transportation mode capacity	0.014	0.013	0.015	0.012	0.011	0.012	0.011	0.012
Security	0.023	0.015	0.018	0.017	0.014	0.017	0.012	0.012
Intermodal mode connectivity	0.023	0.017	0.017	0.015	0.014	0.014	0.014	0.013
Geopolitics	0.015	0.011	0.012	0.011	0.012	0.011	0.012	0.010
Environment	0.014	0.010	0.015	0.015	0.012	0.015	0.011	0.012
Transportation cost	0.005	0.023	0.024	0.026	0.023	0.024	0.022	0.023
Transportation time	0.026	0.014	0.014	0.010	0.016	0.016	0.018	0.017
Priority	**0.142**	**0.120**	**0.133**	**0.123**	**0.119**	**0.132**	**0.119**	**0.112**
Rank	**1**	**5**	**2**	**4**	**7**	**3**	**6**	**8**

and Ocean) took 20 days to complete per journey. In contrast, with respect to transportation cost, the cheapest way is route 4, while the most expensive alternative is airway.

To be more detail, with the respect to security factor, it is easy to understand that the airway route took advantage, followed by route 3 (Rail + Ocean through Shanghai port) and route 7 (road way). Meanwhile, with respect to transport mode capacity, the first rank belonged to route 3. It is clear that Shanghai port is the busiest port all over the world in 2019. Hence, the capacity through Shanghai port is bigger than other ports. Furthermore, with respect to security, airway obviously ranked the first place, while road way alternative and the combine of railway and road way alternative ranked the last position. The respect to intermodal mode connectivity and geopolitics showed the same trend with the highest position belonged to route 1 and the lowest weight belonged to route 8. Finally, with the respect to environment factor, the combine of railway and ocean through Shenzhen port is the most friendly and green choice with environment. Meanwhile, the most unfriendly choice with environment belonged to the combine of road and ocean through Shanghai port.

5 Conclusion

In order to increase international trade between China and Vietnam, especially between the southwest of China and the northeast of Vietnam, analyzing transportation routes is becoming a necessary requirement, whether transporting unfinished or finished cargoes. The issue of increasing the international flows among two countries is not only due to the demand of goods but also the demand of rising GVCs.

In this paper, based on the Delphi and CFPR methods, 8 main criteria and 25 sub-criteria influencing electronic component transport were identified and 8 Sino-Vietnam multi-modal transportation routes were evaluated to guide the logistics stakeholders to make rational route choice.

In the current study, the experimental study only focuses on the multimodal transport routes between Chongqing, China and Hai Phong, Vietnam. Furthermore, the data was collected from a small scope. In the future research, the scope of research can be expanded to involve more areas and more commodities. This allows for a comprehensive assessment of the transport routes/networks between Vietnam and China.

Acknowledgements (1) EU-Asia Research Network on Integration of Global and Local Agri-Food Supply Chains Towards Sustainable Food Security (AMD no.777742-56). (2) GCRF (GCRF Covid19 02 FET, funded by Liverpool John Moores University, UK). (3) The China Postdoctoral Science Foundation Funded Project (2019M661085). (4) The Fundamental Research Funds for the Central Universities (3132020237). (5) The Ministry of Education of Humanities and Social Science Project (18YJC630061). (6) Natural Science Foundation of Liaoning Province (2020-BS-068).

References

1. Vietnam Custom Report 2019
2. The World Bank, Global Value Chain Development report 2017, Measuring and analyzing the impact of GVCs on economic development (2017)
3. M.J. Roorda, R. Cavalcante, S. McCabe, H. Kwan, A conceptual framework for agent-based modeling of logistics services. Transp. Res. Part E: Logist. Transp. Rev. **46**(1), 18–31 (2010)
4. B. Ayar, H. Yaman, An intermodal multi commodity routing problem with scheduled service. Comput. Optim. Appl. **53**, 131–153 (2012)
5. J.H. Cho, H.S. Kim, H.R. Choi, An intermodal transport network planning algorithm using dynamic programming intermodal tray: from Busan to Rotterdam in intermodal freight routing. Appl. Intell. **36**, 529–541 (2012)
6. Y. Xie, W. Lu, W. Wang, L. Quadrifoglio, A multimodal location and routing model for hazardous materials transportation. J. Hazard. Mater. **227–228**, 135–141 (2012)
7. J.H. Bookbinder, N.S. Fox, Intermodal routing of Canada. Transp. Res. Part E: Logist. Transp. Rev. **34**(4), 289–303 (1998)
8. X. Yang, J.M.W. Low, L.C. Tang, Analysis of intermodal freight from China to Indian Ocean: a goal programming approach. J. Transp. Geogr. **19**(4), 515–527 (2011)
9. A. Samimi, K. Kawamura, A. Mohammadian, A behavioral analysis of freight mode choice decisions. Transp. Plan. Technol. **34**(8), 857–869 (2011)
10. S.H. Woo, S.N. Kim, D.W. Kwak, S. Pettet, A. Bereford, Multimodal route choice in maritime transportation: the case of Korean auto-part exporter. Marit. Policy Manag. **45**(1), 19–33 (2018)
11. M.B. Regmi, S. Hanaoka, Assessment of intermodal transport corridors: cases from north-east and central asia. Res. Transp. Bus. Manag. **5**, 27–37 (2012)
12. O. Norojono, W. Young, A stated preference freight mode choice mode. Transp. Plan. Technol. **26**(2), 195–212 (2003)
13. D.S. Moon, D.J. Kim, E.K. Lee, A study on competitiveness of sea transport by comparing international transport routes between Korea and EU. Asian J. Shipp. Logist. **31**(1), 1–20 (2015)
14. Y. Wang, G.T. Yeo, A study on international multimodal transport networks from Korea to central asia: focus on second hand vehicles. Asian J. Shipp. Logist. **32**(1), 41–47 (2016)
15. Y. Wang, G.T. Yeo, Intermodal route selection for cargo transportation from Korea to central asia by adopting Fuzzy Delphi and Fuzzy ELECTRE I methods. Marit. Policy Manage. **45**(1), 3–18 (2018)
16. H.H. Olafo, Analysis of the future: the Delphi method. Santa Monica, CA: RAND Corporation (1967)
17. J.C. Brancheau, B.D. Janz, J.C. Werherbe, Key issues in information systems management: 1994–1995 SIM Delphi Results. MIS Q. **20**(2), 225–242 (1996)
18. S. Hayne, C. Pollard, A comparative analysis of critical issues facing Canadian information systems personnel: a national and global perspective. Inform. Manage. **38**(2), 73–86 (2000)
19. V.S. Lai, W. Chung, Managing international data communications. Commun. ACM **45**(3), 89–93 (2002)
20. R. Loo, The Delphi method: a powerful tool for strategic management. Policing: Int. J. Police Strategies Manage. **25**(4), 762–769 (2002)
21. M.R. Geist, Using the Delphi method to engage stakeholders: a comparison of two studies. Eval. Program Plan. **33**, 147–154 (2010)
22. P.-F. Hsu, H.-Y. Chiang, C.-M. Wang, Optimal selection of international exhibition agency by using the delphi method and AHP. J. Inform. Optimization Sci. **32**(6), 1353–1369 (2011)
23. C. Okoli, S.D. Pawlowski, The Delphi method as a research tool: an example, design considerations and applications. Inform. Manage. **42**(1), 15–29 (2004)
24. A.L. Delbecq, A.H. Van de Ven, D.H. Gútafson, Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes, Scott Foresman and Company: Minneapolis, MN, USA (1975)
25. E. Herrera-Viedma, F. Herrera, F. Chiclana, M. Luque, Some issues on consistency of fuzzy preference relations. European J. Operat. Res. **154**(1), 98–109 (2004)

26. Y.H. Chen, R.J. Chao, Supplier selection using consistent fuzzy preference relations. Expert Syst. Appl. **39**, 3233–3240 (2012)

Integrated Planning for the Intermodal Container Terminals of the CR Express based on Markov Process

Cuijie Diao, Shujuan Guo, Gang Li, Xiaohan Wang, and Zhihong Jin

Abstract The Belt and Road initiative (BRI) is committed to strengthening the connectivity among countries and building a comprehensive transport network. The China Railway Express (CR Express) as an important part of the BRI has become a new type of intercontinental trade transport mode between Asia and Europe. With the development of the CR Express, China's coastal ports have become the transfer hub connecting the CR Express with other Asian countries. In this paper, an intermodal container terminal system composed of a maritime container terminal and a railway container terminal is studied. The system consists of railway and truck as two kinds of distribution modes, which is considered as the dual-channel supply chain system. This paper explores the Markov process used in that supply chain system to study the intermodal container terminal system. Numerical analysis manifests the influence of the railway distribution rate. The result shows that cooperation strategy can help to make better use of the reserved terminal and reduce the expected storage time. The cooperation strategy is beneficial to both the terminal operators and the consigners.

Keywords Intermodal container terminal · Markov process · Total revenue · Expected storage time

C. Diao · S. Guo · G. Li · X. Wang · Z. Jin (✉)
College of Transportation Engineering, Dalian Maritime University, Dalian, China
e-mail: jinzhihong@dlmu.edu.cn

C. Diao
e-mail: dcj0601@dlmu.edu.cn

S. Guo
e-mail: guoshujuan@dlmu.edu.cn

X. Wang
e-mail: 1002924430@qq.com

G. Li
School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, China
e-mail: ligangpe2012@hotmail.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_35

1 Introduction

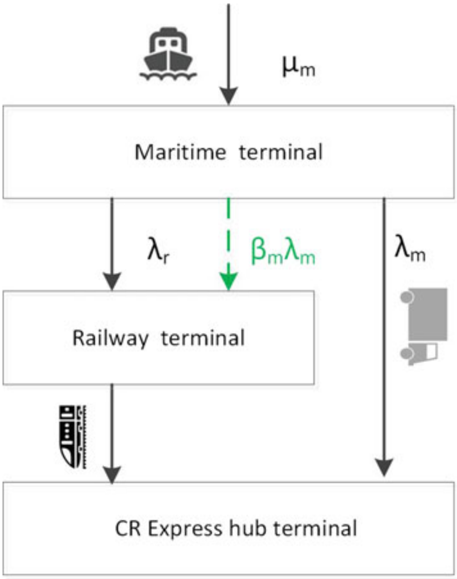
The construction of the strategies of the Belt and Road initiative (BRI) will build a comprehensive transport network and promote the economic development. China Railway Express (CR Express) is an important part of the BRI. With the development of the BRI, the influence of CR Express has gradually extended to Southeast Asia, Japan and South Korea. Coastal ports have become an important node between CR Express and other Asian countries.

This paper studies the intermodal container terminals system in coastal areas is shown in Fig. 1. It is composed by the maritime container terminal and the railway container terminal. The containers will transport to CR Express terminal by train or by truck. Both of the terminals have reserved some storage yard for the CR Express.

The operation of intermodal terminal system is as followed. The system of intermodal terminals will transport containers that arrived at coastal port by sea from maritime terminal to hub terminal of CR Express. The containers arrived at the system are unloaded from the ship in accordance with a Poisson process at constant rates μ_m.The transportation demand by truck arrives at system according to a Poisson process at constant rate λ_m which are satisfied one by one, while the transportation demand by train arrives at system in according to a Poisson process at constant rate λ_r. The railway terminal provides batch service and adopts the fixed-length accumulation mode. The train only departs, when the number of containers in the railway terminal has reached the full capacity of the train.

The system of the intermodal terminals has two levels inventories. The containers in the railway terminal transported by railway are transferred from maritime terminal

Fig. 1 The operation of the intermodal terminal system



according to a Poisson process at constant rates λ_r. It is same as the distribution of the transportation demand by train. The intermodal container terminals system is similar to two-echelon dual-channel supply chain system. Therefore, the Markov analysis is explored in this paper to calculate the revenue.

The contributions of this paper to the literatures are as follows. This paper uses the Markov process to model and analyze the intermodal container terminals for CR Express, fewer scholars take terminal as the research object and consider using the Markov analysis method. Firstly, this paper analyzes the impact of railway distribution rate on the total revenue of the intermodal terminal system. Then, the paper assumes that the maritime terminal and the railway terminal can share the information of the volume of containers in the terminal with each other, and evaluates the cooperation strategy of the intermodal transportation system under different demand rate and arrival rate.

The rest of this paper is organized as follows. The literature review is in Sect. 2. In Sect. 3, an intermodal transportation system model based on Markov process is described including the model assumptions, the operation policy, the Markov process model and the performance measures. In Sect. 4, the numerical analysis is described. Finally, in Sect. 5, the key points are summarized and future research work is mentioned.

2 Literature Review

Many studies have focused on the location of the hub terminal in the future transportation network of CR Express. Wang et al. found the CR Express has its economic transport hinterlands, it is essential to establish a hub-and-spoke transport network and build the transport hubs [1]. From the perspective of freight costs, Jiang et al. applied binary Logit model to compare 5 typical routes of CR Express routes and analyzed the hinterland pattern of CR Express routes. The results show that Chongqing is more likely to become a regional hub of CR Express [2]. Wei et al. studied a logistics network connecting the inland regions by dry ports based on a two-stage logistical gravity model. Dry ports play an important role in the multimodal transport centering on the BRI. Therefore, it is found that the hubs for CR Express will gradually be formed [3].

Some scholars also studied the change of international transport pattern under the influence of the BRI, including the cooperation between the BRI and the original ocean shipping. Wang and Yeo obtained the route from Korea to Central Asia under the BRI with integrated Fuzzy Delphi and Fuzzy methods. The results show that Incheon to Qingdao to Horgos to Almaty is preferred [4]. Sun et al. devised double auction mechanisms for the intermodal transportation service procurement problem under the influence of the BRI and provided several managerial implications [5]. Yang et.al established a bi-level programming model to reconstruct the shipping service network between Asia and Europe by considering the New Eurasian Land Bridge rail services and Budapest-Piraeus railway [6].

Through existing literature, Chan et al. found there is a lack of studies on the BRI from the perspective of logistics and supply chain management [7]. Sheu and Kundu focused on the dynamic and stochastic problems brought by integration of the BRI and international logistic network. A spatial-temporal interaction model combined with Markov chain is used to forecast time-varying logistic distribution flows [8]. The Markov process can also be used to analyze the intermodal terminals system. The model applied in this paper is based on the two-echelon and dual-channel supply chain model proposed by Takahashi et al. which was originally used in the inventory control policy considering production and delivery costs [9].

Some scholars have applied the queuing theory to the research of container truck reservation problem, railway system evaluation, container truck passing capacity at wharf gate and so on. Previous studies have found that the arrival rate of ships and the arrival of transport demand are approximately subject to Poisson distribution. The Markov chain method is suitable for analyzing systems subject to a specific distribution. In this paper, a model of intermodal terminal system is established. The total revenue has been taken as the performance measure, the performance measure is analyzed based on Markov process. This paper considers the two situations of cooperation and non-cooperation and puts forward a reasonable operation strategy.

3 The Model of Intermodal Terminal System

3.1 The Markov Process of Intermodal Terminal System

The Markov process is shown in Fig. 2.The parameters in the figure are explained as follows. S_m and S_r are the maximum inventories of maritime container terminal and railway container terminal reserved for CR Express, respectively. The loading time of train is according to an exponential distribution with mean μ_l.When the number of containers accumulated in the railway terminal reaches the capacity of the train, the containers are loaded to the train and the train will depart.

In the example shown in Fig. 2, $S_m = 9$, $S_r = 6$ and $c = 4$. c is the full capacity of a train. The circles indicate the system state (x, y), x is the number of containers at the maritime container terminal, and y is the number of containers at the railway container terminal. The explanations of the transition rates in Fig. 2 are as follows. If one container picked up from maritime terminal to CR Express hub terminal by truck at rate λ_m, the state (x, y) changes to state $(x - 1, y)$. If one container from the ship is unloaded to the maritime terminal at rate μ_m, the state (x, y) changes to state $(x + 1, y)$. If one container is transported from maritime terminal to the railway terminal at rate λ_r, the state (x, y) changes to state $(x - 1, y + 1)$.

There are three scenarios considering the accumulation and transfer state of the system. They are accumulation, acceleration of accumulation scenario and train departure scenarios. The three scenarios are described as follows.

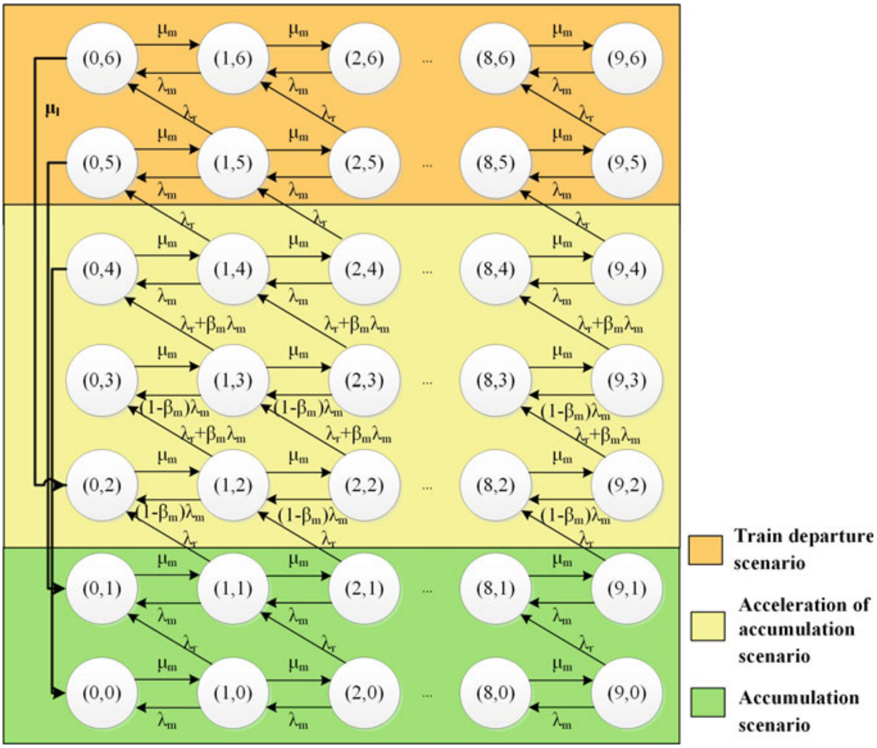


Fig. 2 Markov process of the intermodal terminal system

- Accumulation scenario: If $y < c/2$, the containers are normally accumulated in the railway terminal, the container transported from the maritime terminal to the railway terminal at rate λ_r.
- Acceleration of accumulation scenario: If $c/2 \leq y \leq c$, it means that the train is near to departure. We consider that the maritime terminal and the railway terminal share information with each other. The carriers may notice the train will departure soon. Some proportion of the containers previously planning to transport by truck switch to choose train because of the waiting time will be decreased dramatically. The transfer rate is defined as β_m. The state (x, y) changes to state $(x-1, y)$ at rate $(1-\beta_m)\lambda_m$ in this scenario.
- Train departure scenario: If $y \geq c$, there is enough inventory at the railway terminal to ran a train. The containers are loaded to a train at rate μ_l and the state (x, y) changes to state $(x, y-c)$.

3.2 The Generator Matrix of Intermodal Terminal System

To represent the system transitions in a compact form, we arrange the states of the system in the increasing lexicographic order and obtain the generator matrix. The state transitions between systems expressed in matrix form, the general matrix *P* as follows.

$$P=\begin{array}{c|ccccccccccc|}
 & (x,0) & (x,1) & \cdots & (x,c/2) & (x,c/2+1) & \cdots & (x,c) & (x,c+1) & \cdots & (x,S_y-1) & (x,S_y) \\
\hline
(x,0) & A-D_1 & C & & & & & & & & & \\
(x,1) & & A-D_1 & \ddots & & & & & & & & \\
\vdots & & & \ddots & & & & & & & & \\
(x,c/2) & & & & A & C+D & & & & & & \\
(x,c/2+1) & & & & & A & \ddots & & & & & \\
\vdots & & & & & & \ddots & & & & & \\
(x,c) & B & & & & & & A-B_0-D_0 & C & & & \\
(x,c+1) & & B & & & & & & A-B_0-D_0 & \ddots & & \\
\vdots & & & \ddots & & & & & & \ddots & & \\
(x,S_y-1) & & & & & & & & & & A-B_0-D_0 & C \\
(x,S_y) & & & & & & & & & & & A_0-B_0
\end{array}$$

We have matrices *A*, A_0, *B*, B_0, *C*, D_0 and *D* as follows.

$$A=\begin{array}{c|cccccc|}
 & (0,y) & (1,y) & (2,y) & \cdots & (S_m-1,y) & (S_m,y) \\
\hline
(0,y) & -\mu_m & \mu_m & & & & \\
(1,y) & \lambda_m & -(\mu_m+(1-\beta_m)\lambda_m+\lambda_y) & \mu_m & & & \\
(2,y) & & \lambda_d & -(\mu_m+(1-\beta_m)\lambda_m+\lambda_y) & \ddots & & \\
\vdots & & & \ddots & \ddots & & \\
(S_m-1,y) & & & & & -(\mu_m+(1-\beta_m)\lambda_m+\lambda_y) & \mu_m \\
(S_m,y) & & & & & \lambda_d & -((1-\beta_m)\lambda_m+\lambda_y)
\end{array}$$

$$D=\begin{array}{c|cccccc|}
 & (0,y+1) & (1,y+1) & (2,y+1) & \cdots & (S_m-1,y+1) & (S_m,y+1) \\
\hline
(0,y) & 0 & & & & & \\
(1,y) & \beta_m\lambda_m & 0 & & & & \\
(2,y) & & \beta_m\lambda_m & \ddots & & & \\
\vdots & & & \ddots & \ddots & & \\
(S_m-1,y) & & & & \ddots & 0 & \\
(S_m,y) & & & & & \beta_m\lambda_m & 0
\end{array}$$

$$D_0=\begin{array}{c|cccccc|}
 & (0,y) & (1,y) & (2,y) & \cdots & (S_m-1,y) & (S_m,y) \\
\hline
(0,y) & \beta_m\lambda_m & & & & & \\
(1,y) & & \beta_m\lambda_m & & & & \\
(2,y) & & & \beta_m\lambda_m & & & \\
\vdots & & & & \ddots & & \\
(S_m-1,y) & & & & & \beta_m\lambda_m & \\
(S_m,y) & & & & & & \beta_m\lambda_m
\end{array}$$

$$B=\begin{array}{c|cccccc|}
 & (0,y-c) & (1,y-c) & (2,y-c) & \cdots & (S_m-1,y-c) & (S_m,y-c) \\
\hline
(0,y) & \mu_l & & & & & \\
(1,y) & & \mu_l & & & & \\
(2,y) & & & \mu_l & & & \\
\vdots & & & & \ddots & & \\
(S_m-1,y) & & & & & \mu_l & \\
(S_m,y) & & & & & & \mu_l
\end{array}$$

$$B_0=\begin{array}{c|cccccc|}
 & (0,y) & (1,y) & (2,y) & \cdots & (S_m-1,y) & (S_m,y) \\
\hline
(0,y) & \mu_l & & & & & \\
(1,y) & & \mu_l & & & & \\
(2,y) & & & \mu_l & & & \\
\vdots & & & & \ddots & & \\
(S_m-1,y) & & & & & \mu_l & \\
(S_m,y) & & & & & & \mu_l
\end{array}$$

$$C = \begin{array}{c|ccccccc|} & (0,y{+}1) & (1,y{+}1) & (2,y{+}1) & \cdots & (S_m-2,y{+}1) & (S_m-1,y{+}1) & (S_m,y{+}1) \\ \hline (0,y) & 0 & & & & & & \\ (1,y) & \lambda_r & 0 & & & & & \\ (2,y) & & \lambda_r & 0 & & & & \\ \vdots & & & \ddots & \ddots & & & \\ (S_m-2,y) & & & & \ddots & 0 & & \\ (S_m-1,y) & & & & & \lambda_r & 0 & \\ (S_m,y) & & & & & & \lambda_r & 0 \end{array}$$

Let π_{xy} be the steady-state probability of the state with the number of containers at the maritime container terminal *x*, the number of containers at the railway container terminal *y*. The steady state probability π is defined as $\pi = [\vec{\pi}_0, \vec{\pi}_1, \ldots, \vec{\pi}_n, \ldots, \vec{\pi}_{S_y}]$, where $\vec{\pi}_n = [\pi_{0n}, \pi_{1n}, \pi_{2n}, \ldots, \pi_{S_m-1,n}, \pi_{S_m,n}]$, $\forall n \geq 0$. Based the standard solution method to obtain the steady state probability of Markov process, the steady state probability matrix π obtained by solving the equations $\pi \cdot P = 0$ and subject to the constraint that the sum of the all probability of matrix π is equal to 1. In this paper, GTH algorithm is used to get the steady-state probability by solving the $\pi P = 0$ and $\pi e = 1$.

3.3 Performance Measure

The main performance measure used in this paper is the total revenue of the maritime terminal and railway terminal. It consists of the container operation revenue, the transfer revenue, and the train operation revenue. The revenue is calculated with the steady-state probability $\vec{\pi}_n$. Let *i* present the terminals. $i = 1$ represents maritime terminal and $i = 2$ means railway terminal.

Let tsq_i be the container storage quantity at terminal *i*. tsq_i is calculated by (1).

$$tsq_1 = \sum_{x=0}^{S_m} \sum_{y=0}^{S_r} x\pi_{xy}, \; tsq_2 = \sum_{x=0}^{S_m} \sum_{y=0}^{S_r} y\pi_{xy} \tag{1}$$

Let CST_i be the unit container operation revenue at terminal *i*. Ctr_i is the container operation revenue at terminal *i* is calculated using

$$Ctr_i = CST_i \times tsq_i \tag{2}$$

Let *ttq* be the container transfer quantity at the port to the railway terminal. *ttq* is obtained by (3).

$$ttq = \sum_{x=0}^{S_m} \sum_{y=c/2}^{c} \beta_m \lambda_m \times \pi_{xy} \tag{3}$$

Let *CSP* be the unit container transfer revenue at the railway terminal. Container transfer revenue C_{tra} is calculated using

$$C_{tra} = CSP \times ttq \tag{4}$$

Let *vnum* is the number of trains to departure, which is obtained by (5).

$$vnum = \sum_{x=0}^{S_m} \sum_{y=c}^{S_r} \mu_l \pi_{xy} \tag{5}$$

Let *CPR* be the train operation revenue per number of train and the train operation revenue C_{opt} is calculated using

$$C_{opt} = CPR \times vnum \tag{6}$$

Finally, the total revenues under the accumulation modes are calculated using (7) as follows.

$$TC = Ctr_1 + Ctr_2 + C_{tra} + C_{opt} \tag{7}$$

The other performance measures such as expected storage time in the system W, and terminal utilization U, can be calculated using Little's Law. Let W_1 and W_2 be the expected storage time in the system and the railway container terminal, respectively. W_1 and W_2 are obtained by (8)

$$W_1 = (tsq_1 + tsq_2)/\mu_m, \quad W_2 = tsq_2/\lambda_r \tag{8}$$

Let U_1 and U_2 be the utilization of maritime container terminal and the railway container terminal, respectively. U_1 and U_2 are obtained by (9).

$$U_1 = tsq_1/S_m, \quad U_2 = tsq_2/S_r \tag{9}$$

4 Numerical Analysis

In this section, there are three main research questions. The first question is the impact of railway distribution rate α on the total revenue of the system. The second question is the impact of cooperative strategy on the total revenue of the system. The third question is the impact of cooperative strategy on the utilization and waiting time of the system.

The full capacity of the train is $c = 40$. The container transport by truck to railway transfer rate β_m was set 0.5. The capacities of maritime terminal and railway terminal are $S_m = S_r = 60$.

The cost parameters considered in this paper are set as follows. The loading time of containers on the train follows an exponential distribution with $\mu_l = 5$. The unit container operation revenue at maritime and railway terminals are $CST_1 = 5$ and $CST_2 = 5$, respectively. The unit container transfer revenue is $CSP = 15$. The unit train departure revenue is $CPR = 150$.

4.1 Effects of the Railway Distribution Rate

The impact of railway distribution mode preference rate α on the total revenue is studied. α is set to 0.1, 0.2, …, and 1.0. $\alpha = 1$ means all containers are transported by the railway to the CR Express hub terminal and no containers are transported by the truck. In this section, the following parameters are fixed as follows. (μ_m, λ) is set at (15, 13).

Figure 3 shows the total revenues in different value of α. With the increase of the railway transportation preference rate, container operation revenue of maritime is almost the same. The container operation revenue of railway and train departure revenue increase as α increases. The transfer revenue increases first and then decrease as the value of α increases, because of the demand transport by train and the accumulation of containers in the railway terminal increase. In total, the total revenues increase as the value of α increases. Therefore, it is suggested for consigner to choose the railway mode.

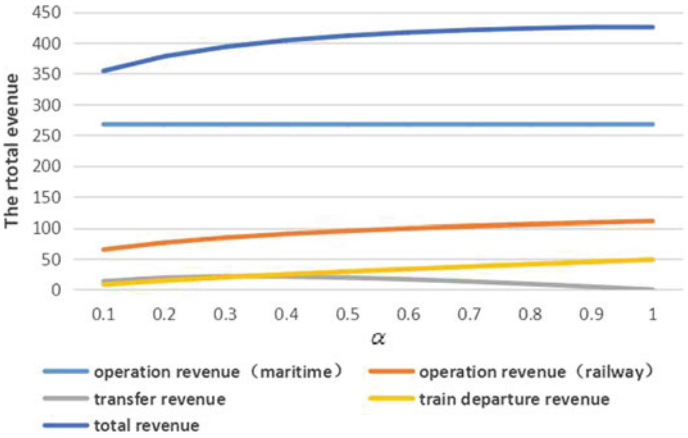


Fig. 3 The total revenue of different transportation mode preference rates

Table 1 The difference of the total revenues between with and without cooperation

(μ_m, λ)	The total revenue	
	With cooperation	Without cooperation
(15, 13)	416.96	402.10
(25, 20)	461.21	434.50
(40, 30)	511.57	467.97
(60, 45)	578.93	510.31

4.2 Total Revenues in the Situation of with Cooperation and Without Cooperation Between Two Terminals

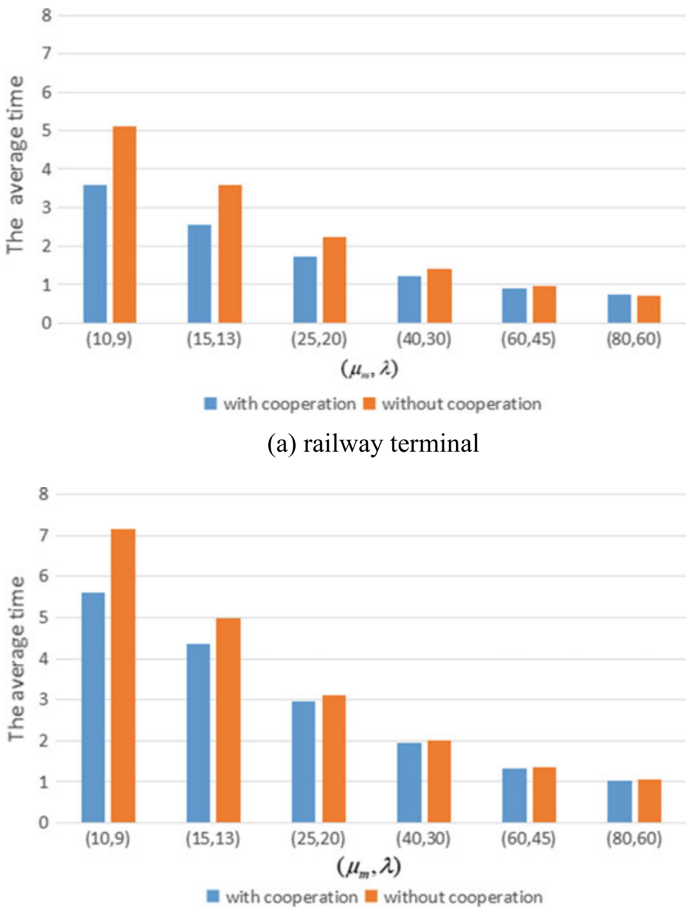
The scenarios of two terminals with and without cooperation under different arrival rate and demand rate are compared. In the strategy without cooperative that the maritime terminal and railway terminal work separately.

It can be seen from Table 1 that the cooperation strategy can generate more total revenue regardless of the transportation demands. However, when the transportation demand is small as (15,13), the revenue difference between with cooperation strategy and without cooperation strategy is small. When the transportation demand is large as (60,45), the revenue difference between with cooperation strategy and without cooperation strategy is large.

4.3 Analysis of the Average Storage Time and the Utilization

The average storage time of different arrival rate μ_m and total demand rate λ is shown in Fig. 4. The average storage time of the railway terminal and the system with cooperation is no more than that without cooperation in all demand cases. The cooperation can help to reduce the waiting times and speed up the turnover. It is beneficial to promote consigner to choose railway transportation, which is more environmental and economic.

The utilization of different arrival rate μ_m and total demand rate λ of the maritime terminal and the railway terminal is shown in Fig. 5. With the increase of the transportation demand, the utilization rate of the maritime terminal increases to near the maximum inventories. Under the situation of cooperation, the intermodal terminal system can be better utilized and the reserved railway terminal can be more utilized.



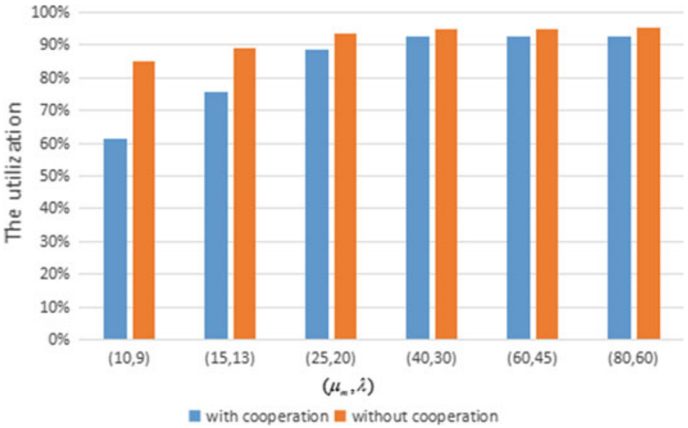
(a) railway terminal

(b) the system

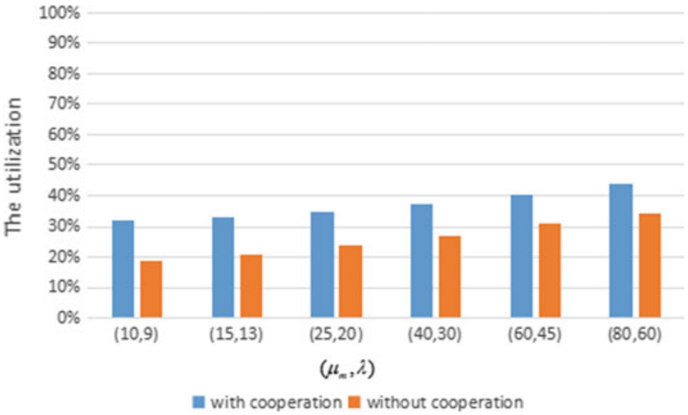
Fig. 4 The average storage time of different arrival rate μ_m and total demand rate λ of railway terminal and the intermodal terminal system

5 Conclusion

In this paper, the model based on Markov process is applied to carry out the impact of railway distribution rate on the total revenue of the intermodal transportation system, and the impact of cooperative strategy on the total revenue and operation of the system under different arrival rate and demand rate. Previously, few scholars have tried this method to the research of intermodal terminal. Numerical analysis shows that the strategy with cooperation will yields higher revenue for both terminals. The cooperation can help to make better use of the reserved terminal and reduce the waiting times of consignees, it suggested for consigner to choose the railway mode.



(a) maritime terminal



(b) railway terminal

Fig. 5 The utilization of different arrival rate μ_m and total demand rate λ of the maritime terminal and the railway terminal

For future research, our work can be extended to include batch arrivals and batch service which are more according with the actual situation of maritime terminal and railway terminal. For the system, different railway accumulation modes can be considered, this paper only considers the fixed-length mode, in fact, most of China's railway terminals use fixed-time mode.

Acknowledgements This research is partially supported by National Natural Science Foundation of China (71572023, 71302085), European Commission Horizon 2020 (MSCA-RISE-777742-56); Leading Talents Support Program of Dalian (2018-573), Project funded by China Postdoctoral Science Foundation (2018M631781), Natural Science Foundation of Liaoning Province, China (20180550262), and Fundamental Research Funds for the Central Universities (3132019301, 3132020301).

References

1. J. Wang, J.J. Jiao, Y. Jing, L. Ma, Transport hinterlands of border ports by China-Europe express trains and hub identification. Prog. Geogr. **36**, 1332–1339 (2017)
2. Y.L. Jiang, J.B. Sheu, Z.X. Peng, B. Yu, Hinterland patterns of China Railway (CR) express in China under the Belt and Road Initiative: a preliminary analysis. Transp. Res. Part E Logistics Transp. Rev. **119**, 189–201 (2018)
3. H.R. Wei, Z.H. Sheng, P.T. Lee, The role of dry port in hub-and-spoke network under belt and road initiative. Marit. Policy Manage. **45**, 370–387 (2017)
4. Y. Wang, G. Yeo, Intermodal route selection for cargo transportation from Korea to Central Asia by adopting Fuzzy Delphi and Fuzzy ELECTRE I methods. Marit. Policy Manage. **45**, 3–18 (2017)
5. J. Sun, G. Li, S.X. Xu, W. Dai, Intermodal transportation service procurement with transaction costs under belt and road initiative. Transp. Res. Part E Logistics Transp. Rev. **127**, 31–48 (2019)
6. D. Yang, K. Pan, S.A. Wang, On service network improvement for shipping lines under the one belt one road initiative of China. Transp. Res. Part E Logistics Transp. Rev. **117**, 82–95 (2018)
7. H.K. Chan, J. Dai, X.J. Wang, E. Lacka, Logistics and supply chain innovation in the context of the Belt and Road Initiative (BRI). Transp. Res. Part E Logistics Transp. Rev. **132**, 51–56 (2019)
8. J.B. Sheu, T. Kundu, Forecasting time-varying logistics distribution flows in the One Belt-One Road strategic context. Transp. Res. Part E Logistics Transp. Rev. **117**, 5–12 (2017)
9. T. Katsuhiko, T. Aoi, D. Hirotani, K. Morikawa, Inventory control in a two-echelon dual-channel supply chain with setup of production and delivery. Int. J. Prod. Econ. **133**, 403–415 (2011)

Research on Quantitative Stock Selection Method Based on Random Forest

Haining Yang and Xuedong Gao

Abstract Stock investment has always been a hot issue in the financial and investment fields. The research of stock quantitative investment method based on modern data mining method and machine learning technology is a brand new field. Random forest is a prediction method integrating multiple decision trees. This paper studies the application of random forest in the quantitative stock selection of stocks, selects the annual report data of China and Shenzhen 300 constituent stocks from 2014 to 2018, and compares the prediction of stock investment returns by using decision tree and random forest method respectively. The applicability and prediction performance of the random forest method are verified, and the importance and interrelationship of annual report indicators are preliminary explored.

Keywords Decision tree · Random forest · Quantitative stock selection

1 Introduction

Quantitative investment is an important way of stock investment, which refers to the process of using computer technology and specific mathematical model to perform the investment strategy according to the investment concept. Quantitative investment originated in the 1950s. Its concept was first written by Frank J. Fabozzi, an American scholar, in his "challenges in quantitative equity" professor Fabozzi wrote "The behavior of managing assets through information and personal judgment is called fundamental investment or traditional investment, and the behavior of generating investment decision by computer model according to fixed rules is called quantitative investment." In 1971, Barclays Global Investors released the world's first passively managed index fund, marking the official start of quantitative investment.

H. Yang (✉) · X. Gao
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: yanghaining@apiins.com

X. Gao
e-mail: gaoxuedong@manage.ustb.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_36

Compared with the traditional qualitative investment method, the quantitative investment method has certain fixity and regularity. It can make investment decision by establishing mathematical model, and finally achieve efficient and accurate stock selection. It can solve the problem that traditional investment methods are difficult to make accurate decision-making judgments when facing high-dimensional and massive data, and can effectively avoid the subjectivity of decision-making caused by qualitative investment. Quantitative investment methods are relatively mature in foreign countries. According to the survey, more than 30% of investment products in the U.S. stock exchange market use quantitative investment technology.

The nature of quantitative stock selection is a classification problem in the field of data mining. One way to quantify stock selection is to build a stock selection index using technical indicators as guidance. At present, many scholars in China have explored in terms of quantitative stock selection. For example, Xie and Hu constructed a multi-factor quantitative stock selection model and tested it with CSI 300 constituent stocks, and finally found that the quantitative investment strategy constructed by them can achieve higher excess returns [1]. Li and Lin [2] designed a quantitative investment algorithm based on machine learning and technical indicators, using technical indicators as input variables, and using different machine learning algorithms to predict the direction of the stock's rise and fall after several days, then use the prediction to build a portfolio.

Quantitative stock selection is to use a quantitative method to build a stock portfolio and make the stock portfolio return higher than the benchmark return. The biggest advantage of the quantitative stock selection method is that it can exclude the influence of investors' subjective factors. The construction of quantitative stock selection models is generally based on two aspects: one is fundamental analysis, and the other is technical analysis. Fundamental analysis refers to quantitative analysis based on fundamentals, rather than traditional fundamental analysis based on qualitative analysis. The types of stock selection models constructed based on fundamentals are more diversified, and when measuring the value of stock investments, they will focus on the operating and financial conditions of listed companies and their industrial environment. The stock selection method based on technical analysis is to analyze the current stock trend and speculate on the future stock price trend through various technical indicators of stock trading. Machine learning is a new method of quantifying stock selection in recent years, and most of them belong to black box algorithms. The principle of the machine learning algorithm is to build a specific model, use the algorithm to analyze the existing data and obtain the rules, then predict the future trend based on the rules obtained.

2 Related Work

2.1 Overview of Quantitative Stock Selection Methods

Many scholars have used support vector machine methods when researching on quantitative investment using machine learning methods. For example, Chen and Yu designed a stock selection model using a support vector machine based on a heuristic algorithm, and found that the model selected by the model was used to select The stock portfolio is significantly higher than the benchmark return over the same period [3]. Su and Fu used the core principal component genetic algorithm to conduct principal component analysis on the company's financial indicators, and combined the support vector machine method to make long-term and short-term predictions of stocks to assist stock selection [4]. Some scholars have used random forest algorithms to study stock selection. For example, Cao et al. built an index system based on the GARP investment strategy, and used the random forest algorithm to accurately classify stocks [5]. The essence of support vector machine and random forest algorithm is a classification problem. Similar to neural networks, a training set and a test set are also needed. The training set plays a crucial role in the effectiveness of the final model. This type of classification algorithm needs to set learning goals, but for stock data, determining learning goals is a more difficult problem.

2.2 Decision Tree

Data classification is a common problem in daily life. The problem that needs to be solved in many areas is essentially data classification, such as geological image classification, biological information classification, the level of customer classification, etc. Classification has become one of the most important research projects in the field of data processing, the traditional statistical data classification methods include cluster analysis and Bayes classification. In recent years, with the expansion of various data storage capacities, more and more multi-dimensional complex data have been accumulated, and traditional data classification algorithms are not good at processing these multi-dimensional complex data. Therefore, scholars have begun to study and propose multi-dimensional Data classification algorithm.

Decision tree is a method commonly used in the field of data exploration. It mainly uses classification models as the main structure to inductively classify specific data and establish a logical basis for decision analysis. Compared to other data mining algorithms, the preparation of the decision tree in data is often simple or unnecessary, and it can handle both data-type and regular-type attributes at the same time. It is not required that the data must have a single attribute. Excellent performance in rule-oriented fields, capable of making feasible and effective results on large data sources in a relatively short time, and producing easy-to-understand rules. After interpretation, people have the ability to understand the expressions of decision trees.

In the field of academic research and practical applications, decision trees are often used in the marketing, finance and insurance, medicine, and engineering. Because each leaf node of the decision tree is an If-Then judgment, it does not require much calculation. It can be classified regardless of whether the input data is a category scale or a continuous scale. Therefore, many domestic and foreign scholars used the concept is applied to the study of investment finance.

Decision tree is a widely used classification algorithm. It is a tree-like classifier that selects split attributes at each internal node for classification, and each leaf node has the same category of data. When the data to be classified is input, the decision tree is a path from the root node to the leaf node through the continuous identification process, and the category of the leaf node of the path is the category to which the sample to be classified belongs. Decision tree is a simple and fast non-parametric classification method. In general, it has a good classification accuracy. However, when the data is complex or has a lot of noise, the decision tree is prone to excessive nodes or excessive configuration problems. This may cause classification accuracy reduced.

Chen believes that accounting earnings information is important information for corporate stakeholders to measure corporate value, determine corporate stock prices, and evaluate the performance of their regulatory agencies [6]. Attempt to use a decision tree to analyze and find out the company's operating cash flow, interest rate of return and previous period Discretionary accrued profits play a decisive role in affecting its extreme earnings management.

Sun et al. using the Synthetic minority oversampling technique Unbalanced corporate credit evaluation DT set model and a Bagging ensemble learning algorithm with dynamic super resolution, named DTE-SBD (Decision Tree Integration), which constructs an effective decision tree integration model for corporate credit assessment, as an important tool for risk management of banks, enterprises and general investors [7].

Malliaris and Malliaris used decision tree technology to combine traditional financial variables such as stock returns, stock volatility, oil prices, the euro exchange rate, and the emerging Cleveland Financial Stress Index to analyze gold price fluctuations, and the resulting classification model can effectively explain Changes in the price of gold [8]. In the study of Caoa et al., artificial neural networks were used to predict stock price changes of companies trading on the Shanghai Stock Exchange [9]. Compare the predictive power of linear models in literature with univariate and multivariate neural network models. The results show that the performance of the neural network is better than the linear model and can effectively predict the performance of stocks in emerging markets.

As for stock technical analysis and reference news, some scholars use decision trees for research. For example, Kamble analyzes moving average convergence/divergence (MACD), relative strength index (RSI), stochastic oscillator (KD), and Bollinger Bands (BB), And found that the use of appropriate processing technology and machine learning models can improve the accuracy of short-term trend prediction [10]. If combined with fundamental analysis and technical data, long-term stock prince prediction is possible. And Al-Nasseria et al. based on more and more evidence that posts on online stock forums will affect stock prices and change investors' investment decisions, and based on this, combine text mining technology, feature selection and decision-making. Tree algorithm is used to analyze and extract semantic terms expressing specific emotions (sell, buy or hold) [11]. An intelligent transaction support system based on sentiment prediction is proposed. In this paper, feature selection filtering method is used to identify the most relevant terms in this article are published, and then a decision tree model is constructed to determine those terms or how to combine them, which can effectively improve the profitability of trading strategies.

2.3 Random Forest

Random Forest is a machine learning algorithm published by Leo [12]. It combines the Bagging machine learning theory proposed in 1996 with the random subspace method proposed by Ho in 1998. The basic operation way is that each decision tree is a classifier, each node of the decision tree is a weak classifier, a random forest is composed of multiple decision trees, and many weak classifiers are formed into a strong classifier. The final decision result is determined by a majority vote. Random forest has two important parameters, one is the number of candidate feature parameters selected when each node of a single decision tree is split, and the other is the number of decision trees in the random forest. Compared with the traditional decision tree, the random forest algorithm has stronger generalization ability and classification effect. Due to its good performance, the algorithm is widely used in bioinformatics, medical research, business analysis, text exploration, semantic classification, Economics, finance and other practical areas, and achieved good results.

Although random forests have been used in different fields, in the study of exchange rate prediction, the application of random forests is not widespread. The related research and literature of random forests are:

Kumar and Thenmozhi researched the predictability of the stock index movement direction, and researched and forecasted the securities trading of the Standard & Poor's CNX NIFTY market from 2000 to 2005 [13]. Empirical results show that the random forest method is better than neural networks, discriminant analysis, and logit models in predicting the direction of stock market movements.

Montillo and Ling proposed the use of random forests to predict a person's age through facial image analysis, to explain the wide range of practical applications and the potential of advertising, and to provide some random forest parameters, which are the advantages of relatively easy initialization [14]. In no prior model, the parts that highlight human features are classified, and the training time is greatly reduced, while maintaining the high regression accuracy of the entire human development.

Rodriguez-Galiano et al. used random forest machine learning classification to cover the coverage of 14 different land categories in southern Spain, using remote sensing data to monitor, map accuracy, sensitivity, and The size and noise of the data set are used to compare the classification accuracy, and the results show that the random forest model performs better than the single decision tree [15].

Theofilatos et al. studies the performance of machine learning technology transactions and the EUR/USD exchange rate, and integrates five supervised learning classification technologies (K-Nearest Neighbors algorithm, Naïve Bayesian Classifier, Artificial Neural Networks, Support Vector Machines and Random Forest) to make a time-pane movement forecast for the EUR/USD exchange rate [16]. The study found that the trading strategies derived from random forest machine learning techniques clearly outperformed the annualized returns and Sharpe ratio calculations of all other strategies.

Qin et al. used gradient boosted random forest to deal with non-linear trading patterns of stocks, and used higher weighted market indicators to construct trading decisions, that is, signals to buy, sell or hold, The 9 stocks and 1 index are used to measure the performance of trading decisions [17]. The empirical results show that the proposed trading method generates excess returns in the buy and hold strategy.

3 Research Design and Methods

3.1 Research Objects and Data Sources

Since the annual reports of most companies in 2019 have not yet been published, in order to make the sample universal, the time period selected in this article is from 2014 to 2018, and the selected financial indicator data is the annual report data of the CSI 300 constituent stocks.

The definition of high-quality stocks in this article is the stock's return on investment exceeds the returns of the CSI 300 Index and the bank's time deposit interest rate for that year. When calculating the return on investment in this study, in order to simplify the analysis process, transaction costs will not be considered. The investment return rate is calculated using the holding period remuneration method. It is bought at the opening price of the first trading day in April after the annual financial report of the listed company is announced, and sold at the closing price of the last trading day in March, plus the rights issue during the period Calculate the return on investment from the dividend income. The formula for calculating the return on stock investment and the return of the CSI 300 Index is as follows:

$$\text{T-year stock investment return rate} = \frac{\mathrm{C} - \mathrm{O} + \mathrm{D} + \frac{\mathrm{A}}{10} * \mathrm{C}}{O} * 100\%$$

$$\text{CSI 300 Index Return } = \left(\frac{C}{O} - 1\right) * 100\%$$

Data processing is the most basic link for subsequent analysis. Therefore, after the data collection is completed, the following principles are adopted for data processing:

1. Exclude some stocks that cannot calculate the stock return for the current year.
2. Exclude stocks with more missing values in financial indicators.
3. Normalize the financial indicator data.

After data processing, a total of 944 samples were selected in this paper.

3.2 Selection of Research Application Indicators

Based on the previous research results, this article proposes the following stock indicators as research variables.

1. *Business indicators*
   - (a) *Gross profit margin and gross profit margin growth rate*: Gross profit margin, also known as sales gross profit margin, is usually expressed as a percentage, and is a measure of corporate profitability. The higher the company's gross profit margin, the stronger its profitability and its ability to control costs. However, for companies of different industries and sizes, the reference value for comparison of gross profit margin is not high, so this study adds the growth rate of gross profit margin to explore whether it has reference value.
   - (b) *Operating profit rate and operating profit growth rate*: The operating profit rate is an index to measure the operating efficiency of an enterprise, and reflects the ability of an enterprise's operators to make profits from its own business operations. The higher the operating profit rate, the more profitable and profitable the company is on its own business; on the other hand, the lower the profitability, the weaker it is. This study also includes the operating interest growth rate to explore whether the increase or decrease of the ratio affects the company's stock price performance.
   - (c) *Pre-tax net interest rate and Pre-tax net profit growth rate*: The net profit margin before tax is the proportional relationship between the net profit before tax and the net operating income. It can be seen that the company's business operation plus the profitability of non-operating income and expenditure. Or the non-operating income and expenditure have improved, so this study also includes indicators of the growth rate of net profit before taxes.
   - (d) *Net profit after tax and Net profit growth rate after tax*: Net interest rate after tax is the proportional relationship between net profit after tax and net operating income. It can be seen that the company's final gross income after deducting all production costs, interest and taxes. The higher the ratio, the more sufficient funds the company can use in future operations and investments. It also means that the company does not rely on debt to distribute dividends. If the net profit after tax increases, it indicates that the company's actual profit has increased, so this study also includes an indicator of the net profit growth rate after tax.

(e) *Current ratio*: The current ratio refers to the ratio of the company's current assets to its current liabilities, and reflects the short-term debt-servicing capacity of the company. Generally speaking, the higher the current ratio, the stronger the ability to realise the assets of the enterprise and the stronger the short-term debt repayment ability; otherwise, the weaker it is.

(f) *Quick ratio*: Quick ratio refers to the ratio of quick assets to current liabilities. The role is to measure the ability of a company's current assets to be immediately realized for repayment of current liabilities.

(g) *Debt ratio*: The debt ratio refers to the ratio of total debt to total assets, and is an important indicator of the capital structure of a company. The source of capital is composed of shareholder funds and external borrowing. Raising funds through borrowing can play the role of financial leverage, which can help improve the return on investment, and the interest expenses can be deducted, so it also has tax advantages. However, its disadvantage is borrowing. When it is too high, the risk is increased due to leverage, and if the operation is not as expected at the same time, there is a risk of failure. Therefore, the debt ratio can tell whether a company's health is healthy.

(h) *Interest coverage ratio*: The interest protection multiple is used to measure the company's ability to pay interest expenses from net profit before interest and tax. The higher the multiple, the higher the degree of protection of the creditor, that is, the higher the debtor's ability to pay interest.

(i) *Accounts receivable turnover*: Accounts receivable turnover rate refers to the ratio of net sales amount divided by accounts receivable plus bills receivable. It calculates the number of times a year receivables are collected and is used to test the ability of an enterprise to recover receivables. The higher the value of this ratio, the better. Conversely, the longer the funds stay outside, the greater the chance of becoming a bad debt.

(j) *Average collection days*: The average collection days is the result of dividing 365 days by the receivables turnover rate. This number of days shows the average time it takes the company to receive the payment after the product is sold, and it is used to calculate the time it takes for the company to receive the payment. The shorter the number of days, the better, which means that the payment will be recovered quickly, and the probability of becoming a bad debt will be reduced.

(k) *Inventory turnover*: The inventory turnover rate is an indicator used to measure the inventory turnover speed of an enterprise, and it symbolizes the ability of an enterprise to sell goods and its business performance. The higher the ratio, the lower the inventory, and the higher the efficiency of capital utilization. However, when the ratio is too high, it may also mean that the company's external inventory may be insufficient and the sales opportunity will be lost. Conversely, if the inventory turnover rate is lower, it means that the company has too much inventory and its operation is sluggish.

(l) *Circulating rate of fixed assets*: The turnover of fixed assets refers to the result of the company's annual net product sales revenue and the average net value of fixed assets. It is mainly used to reflect the turnover of fixed assets of enterprises, and then to measure the efficiency of fixed assets. The higher the ratio, the higher efficiency of the use of fixed assets. On the contrary, it means that the enterprise's utilization rate of fixed assets is low, which may affect its profitability in the future.

2. Stock trading indicators

   (a) *Volume*: Trading volume is a very important indicator for stock trading. For stocks with good fundamentals in financial report data, if no one wants to trade and the trading volume is sluggish, the stock price has less room for imagination and has limited ups and downs. On the contrary, stocks with large trading volume fluctuate greatly. If you can properly analyze whether the financial report is sound or not, buying stocks with sound physique and stable profitability during dips during the market panic is a good buy for long-term investment. This research mainly analyzes the company's annual financial indicators. Therefore, the volume data is mainly based on the annual turnover of individual stocks on the Taiwan Stock Exchange website.
   (b) *Volume growth rate*: If the stock trading volume grows, it means that the company's fundamentals may improve, profits have improved, and investors have good looks. On the contrary, it means that the fundamentals of the company may be out of order, failing to get the favor of investors, and the stock price is likely to fall.
   (c) *Return on equity*: Return on equity is also known as return on equity, return on equity or return on equity. It is an investment return measure that measures shareholders' equity. It mainly reflects the company's ability to use net assets to generate net profit. It is the most commonly used financial reference for many scholars and general investors.
   (d) *Price-to-book ratio*: The price-to-equity ratio is the price per share divided by the net assets per share, and is usually one of the important indicators of whether the stock price is worth the money. In particular, it is valued when evaluating high-risk industries and companies with a large amount of physical assets. Due to daily changes in stock prices, the net value of stock prices is slightly different from daily.
   (e) *Price-to-earning ratio*: The price-earnings ratio is calculated by dividing the price of each stock market by earnings per share, which links the company's stock price with its ability to produce wealth, which is usually used as an indicator that the stock price is cheap or expensive. Daily price-to-earnings ratios are slightly different due to daily changes in stock prices.

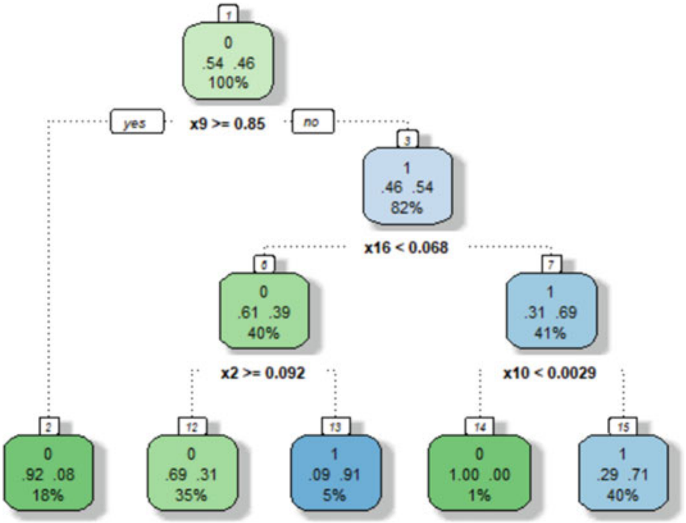


Fig. 1 C5.0 decision tree model

4 Research Results

4.1 Decision Tree Prediction Results

After preprocessing the data, the decision tree is modeled. In this study, the C5.0 algorithm is used to model the data. The model is built by R language programming. The screenshot of the generated C5.0 decision tree model after pruning is shown in Fig. 1 Show. ROC curve and AUC value are often used to evaluate the pros and cons of a binary classifier. Because the classification attributes of the data set used are only two types, the AUC index is used to evaluate the classification performance of the model.

It can be seen from Fig. 2 that the AUC of the decision tree model after pruning is 0.693.

4.2 Random Forest Prediction Results

Random forest is a model composed of many decision trees. The "forest" in a random forest represents a collection of multiple decision trees. The concept of "random" has two meanings: a randomly selected sample and a random subset of interesting features. This article uses the randomForest package in the R language for random forest modeling and the pROC package for visual presentation. The experimental process is as follows:

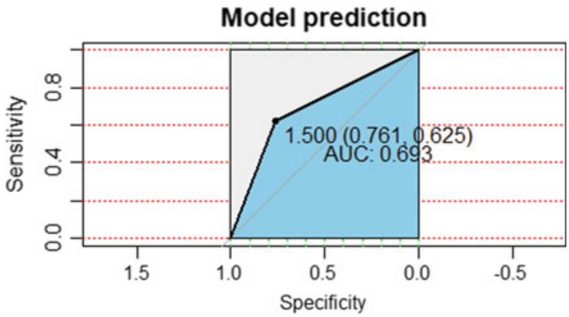


Fig. 2 C5.0 decision tree model ROC Curve

It can be seen from Fig. 3 that the error rate of the model decreases with the increase of trees. When the trees is 100, the error rate remains basically unchanged, so the ntrees parameter of the model is 100. In the process of model parameter optimization, the change of mtry parameter value will also affect the accuracy of the model. For the random forest model, the mtry parameter generally takes the square root of the number of variables. When the ntrees parameter is determined, the model accuracy rate changes with the mtry parameter As shown in Fig. 4.

It can be seen from Fig. 4 that when the mtry parameter value is 7, the model accuracy is the highest, so the mtry parameter value of this model is 7.

It can be seen from Fig. 5 that the AUC value of the final model reaches 0.729.

Compared with the decision tree model after pruning, the random forest model has a certain increase in the AUC value and has achieved good results. This also

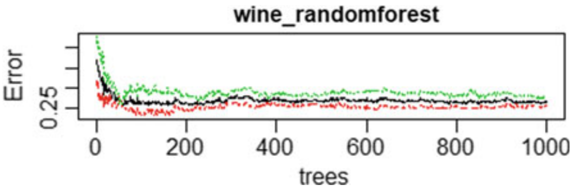


Fig. 3 Error rate under different values of trees

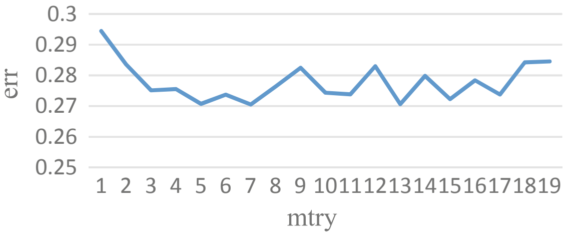
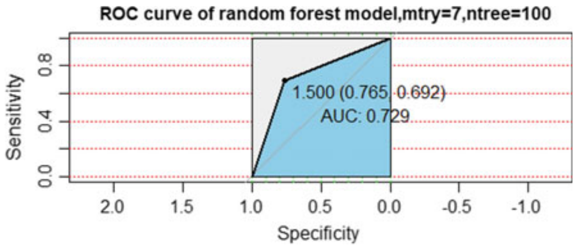


Fig. 4 Error rate under different mtry values

Fig. 5 ROC curve of random forest model



shows that as a "representative method of integrated learning technology," random forest is simple in calculation, easy to implement, and shows strong performance in many real-world tasks. It is one of the mainstream methods for big data analysis and prediction in the future.

From the results of the importance ranking of the Fig. 6 random forest model, the top five attributes of the score include business cycle, PE, PE, PB, ROA, and ROE, reflecting different pairs of attributes. The contribution rate of the model is different, and corporate profitability has a greater impact on the stock's profitability. After hierarchical clustering of these attributes, the results obtained are shown in Table 1. It can be seen from Table 1 that financial indicators of the same nature are grouped together, and there is a strong correlation between different financial indicators.

The clustering results from the above levels also confirm the role of different indicators from another perspective, which has not been deepened in this study and needs further research in the future.

Fig. 6 Attribute Importance

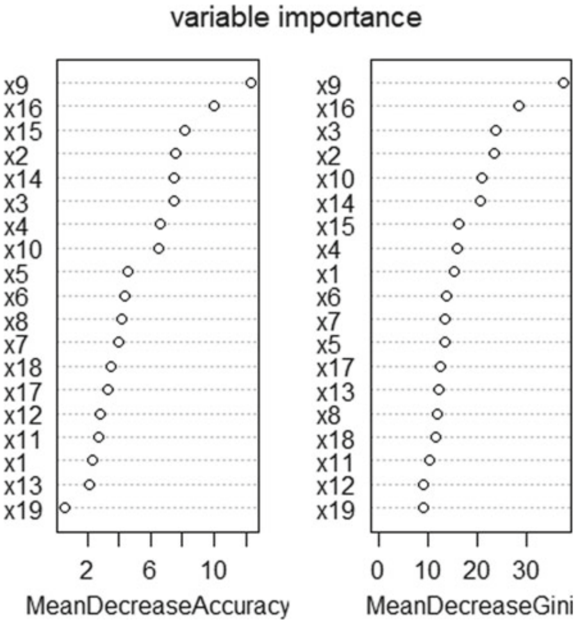


Table 1 Hierarchical clustering results

1	Business cycle
2	Net profit/Total operating income
3	Assets and liabilities
4	Current ratio, quick ratio, PB
5	Accounts receivable turnover (days)
6	ROA, ROA, ROE, ROE
7	Net asset per share BPS, net asset-liability ratio, gross profit, inventory turnover ratio, fixed asset turnover ratio, trading volume, PE ratio, inventory turnover ratio, fixed asset turnover ratio

5 Conclusion

Stock investment is one of the most commonly used options for investment and financial management. How to choose high-quality stocks among many stocks to improve financial returns is a hot topic in current research. In this paper, the financial indexes of some stocks of China Shanghai and Shenzhen 300 Index constituent stocks are selected. Through data preprocessing, a random forest model is constructed, and the quality of the stock is modeled and analyzed.

This article first studies the financial indicator data, selects indicator data that reflects the profitability, development capability, and basic stock situation of the company, performs data preprocessing and indicator screening, and lays a foundation for subsequent modeling and analysis.

In this paper, through comparative analysis of the classification results of the C5.0 decision tree and the random forest algorithm model, it is found that the classification model established by the random forest is more accurate, especially in the face of a small number of samples, its classification accuracy is compared to the single decision tree model A lot higher, this finding has great significance for the classification of imbalanced data.

By ranking the importance of random forest attributes and clustering attribute levels, we found that different financial indicators have different effects on stock quality, and there is a strong correlation between financial indicators. How to measure this correlation has become the direction of future research.

References

1. H.L. Xie, D. Hu, The application of multi-factor quantitative model in investment portfolio—comparative study based on LASSO and elastic net. Stat. Inf. Forum **32**(10), 36–42 (2017)
2. B. Li, Y. Lin, ML-TEA: a set of quantitative investment algorithms based on machine learning and technical analysis. Syst. Eng. Theory Pract. **37**(5), 1089–1100 (2017)
3. R.D. Chen, H.H. Yu, Stock selection model of support vector machine based on heuristic algorithm. Syst. Eng. **32**(2) (2014)

4. Z. Su, X.Y. Fu, Kernel principal component genetic algorithm and improvement of SVR stock selection model. Stat. Res. **30**(5), 54–62 (2013)
5. Z.F. Cao, H. Ji, B.C. Xie, Using random forest algorithm to realize selection of good stocks. J. Capital Univ. Econ. Bus. **16**(2), 21–27 (2014)
6. F.H. Chen, An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree. Soft. Comput. **20**(5), 1945–1960 (2016)
7. J. Sun, J. Lang, H. Fujita, H. Li, Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf. Sci. **425**, 76–91 (2018)
8. A.G. Malliaris, M. Malliaris, What drives gold returns? A decision tree analysis. Financ. Res. Lett. **13**, 45–53 (2015)
9. Q. Caoa, K.B. Leggioa, M.J. Schniederjans, A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. Comput. Oper. Res. **32**, 2499–2512 (2015)
10. R.A. Kamble, Short and long term stock trend prediction using decision tree, in *International Conference on Intelligent Computing and Control Systems (ICICCS)* (2017)
11. A. Al-Nasseria, A. Tuckerb, S. Cesarea, Quantifying StockTwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. Expert Syst. Appl. **42**(23), 9192–9210 (2015)
12. B. Leo, Statistical modeling: the two cultures. Stat. Sci. **16**(3), 199–231 (2001)
13. M. Kumar, M. Thenmozhi, Forecasting stock index movement: a comparison of support vector machines and random forest, in *Indian Institute of Capital Markets 9th Capital Markets Conference Paper* (2006)
14. A. Montillo, H. Ling, Age regression from faces using random forest, in *16th IEEE International Conference* (2009), pp. 2465–2468
15. V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogrammetry Remote Sens. **67**, 93–104 (2012)
16. K. Theofilatos, S. Likothanassis, A. Karathanasopoulos, Modeling and trading the EUR/USD exchange rate using machine learning techniques. ETASR Eng. Technol. Appl. Sci. Res. **2**(5), 269–272 (2012)
17. Q. Qin, Q.G. Wang, J. Li, S.S. Ge, Linear and nonlinear trading models with gradient boosted random forest and application to Singapore stock market. J. Intell. Learn. Syst. Appl. **5**, 1–10 (2013)
18. Z.H. Zhou, *Machine Learning* (Tsinghua University Press, Beijing, 2016), pp. 178–180

MPNC: Improving Reliability of Data Transmission in Wireless Ad Hoc Networks

Mingming Yan, Xu Li, and Wenjun Huang

Abstract Nodes in wireless ad hoc networks usually transmit data via multiple relays, exposing the reliability to risks like multi-hop error accumulation and inter-node interference on the same path. Nowadays, reliable transmission schemes in wireless networks mainly based on redundancy or retransmission, while they can hardly be applied in wireless ad hoc networks directly as they do not consider the uneven link quality. Thus the redundancy may be too large at some hops, so as the repeat times of retransmission, resulting in a great waste of resources. In this paper, we proposed a Multipath Network Coding (MPNC) reliable transmission scheme. Through an analytical comparison with other multipath network coding schemes, we prove that MPNC requires a smaller overhead. Further simulations results show that MPNC has higher reliability than the Forward Error Correction (FEC) and Automatic Repeat Request (ARQ) when the overhead is at the same level.

Keywords Ad hoc networks · Multipath transmission · Network coding · Reliability

1 Introduction

With the characteristics of flexible networking and strong invulnerability, wireless ad hoc networks have received a lot of research attention, and have been widely used in special scenarios such as military and disaster relief. Especially, with the in-depth application of the Internet of Things (IoT) in the field of emergency rescue [1], the reliability of data transmission in wireless ad hoc networks becomes more important [2]. When the existing reliability transmission scheme in the wireless network is

M. Yan (✉) · X. Li · W. Huang
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
e-mail: 18120164@bjtu.edu.cn

X. Li
e-mail: xli@bjtu.edu.cn

W. Huang
e-mail: 16111027@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_37

applied to multi-hop networks, there are many problems such as error accumulation and excessive redundancy due to large fluctuations in the quality of links, which makes it difficult to meet the high-reliability of data transmission requirements in wireless ad hoc networks.

Studies about reliable transmission in wireless ad hoc networks have shown that multipath can reduce the risk of data transmission caused by poor link quality at a single path because multipath scheme uses multiple paths to transmit data between the source node and the destination node [3]. However, the multipath scheme will bring great overhead because of the routing's construction, maintenance, and data transmission in the backup method [4]. Furthermore, the parallel transmission method also has problems due to the disorder of the data packets on receiving. In addition, the network encoding scheme encodes a group of packets into an encoded packet for forwarding and uses redundant encoded packets to recover the damaged data to ensure the reliability of the transmission. The network coding scheme has a high data utilization rate, greater robustness to link quality, and adaptability to topology changes [5]. Therefore, the combination of multipath scheme and network coding scheme can overcome the problems of single link poor stability and error accumulation in multi-hop networks [6]. Xu et al. [7] proposed a multipath routing algorithm based on network coding suitable for satellite networks, which effectively improves the throughput of the network, but the change of the topology of the satellite network is not suitable for wireless ad hoc networks, and the transmission scheme proposed in this paper does not consider the effect of mechanism parameters on performance in the application [8]. Tang et al. [9] proposed a multipath scheme based on network coding with the goal of optimization network throughput. They improved the network throughput and reliability by designing a no-waiting ACK scheme, but they do not analyze the impact on the overhead while improving reliability. Zhang et al. [10] proposed a BigNum Network Coding (BNNC) scheme for vehicle-to-ground multipath communication, which improved the reliability and codec efficiency of vehicular networks. However, when they analyzed the reliability, they do not consider the change of the overhead caused by redundancy, and at the same time, there is no applicability analysis of multi-hop.

In this paper, we proposed a Multipath Network Coding (MPNC) scheme for the reliable transmission in wireless ad hoc networks. The design of MPNC's multipath algorithm is based on the topology maintained by the centralized MAC layer protocol of the wireless ad hoc network, which reduces the routing layer construction and maintenance overhead. After the data packets are encoded, we distribute them to multipath parallel transmission by load balancing strategy. Therefore, the MPNC scheme can reduce overhead and the impact of single link quality fluctuations to improve transmission reliability. Moreover, we established the reliability and overhead mathematical model of the MPNC and compared with the currently commonly used FEC reliability scheme and ARQ reliability scheme [11], we verified the overhead and reliability of the MPNC scheme in multi-hop networks through simulations.

The structure of the paper is organized as follows. The system model is introduced in Sect. 2. In Sect. 3, we introduce the implementation of the MPNC scheme. In

Sect. 4, we introduce the performance mathematical model of the MPNC scheme. In Sect. 5, we verify the performance of the MPNC scheme through simulations. In Sect. 6, we conclude the paper.

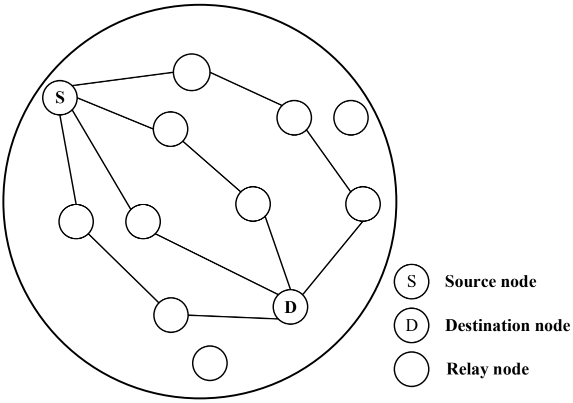
2 System Model

In this section, we briefly introduce the system model of the paper, which mainly includes transmission topology and multipath types.

The multipath transmission model in this paper is shown in Fig. 1. We consider a wireless ad hoc network which node density is ρ, the number of nodes is evenly distributed and the communication radius of the node is r_0. The model consists of a source node S, a destination node D, and several relay nodes. In Fig. 1, the source node encodes the data packets before sending them, and sends the encoded packet in parallel on multipath, the relay node forwards the received encoded packet to the next-hop node, and the destination node will get the original data by decoding the encoded packets after receiving a certain amount packets.

Generally, multipath can be divided into winding multipath and disjoint multipath according to the choice of links and nodes. Multipath transmission can be divided into the parallel transmission and backup transmission according to whether data can be transmitted on multiple paths at the same time. In order to reduce the overhead of the data transmission process and improve the utilization rate of the data packets, we adopt the parallel mode of multipath transmission. In addition, in order to reduce the impact of the common node of different paths due to excessive load or node failure, disjoint multipath is used for transmission in this paper.

Fig. 1 Multipath transmission topology of MPNC scheme



3 The MPNC Scheme Overview

Based on the multipath transmission model shown in Fig. 1, in this section, we briefly introduce the MPNC reliable transmission scheme from three aspects: disjoint multipath routing algorithm, network coding transmission scheme, and dynamic load balancing strategy.

3.1 Multipath Routing Algorithm

Most of the current multipath routing algorithms are improved based on the routing protocol of the network layer, such as changing the process of the node sending and replying RREQ and RREP. In this paper, we propose a multipath routing algorithm based on the centralized MAC layer protocol of the wireless ad hoc network, which reduces the network layer overhead in the process of route establishment and maintenance. The specific algorithm design process is as follows:

The data transmission and resource allocation of the centralized MAC layer protocol is realized by a tree structure, so a scheduling tree structure needs to be maintained, and the scheduling tree is generated by establishing a network-wide connection relationship matrix maintained by base station nodes. Therefore, in the multipath routing algorithm designed in this paper, we will query the next-hop node connected to the current node through the DFS algorithm according to the connection relationship matrix maintained by the base station node. Then we will start from the next-hop node and traverse to find the node connected to it until the query to the destination node. In this process, we set the flag bit for the nodes that have appeared in the query to avoid the occurrence of duplicate nodes in the searched path. In the process of sending and receiving data in the centralized MAC layer protocol, the child node needs to send a resource request to the base station node, and then the base station node performs scheduling and resource allocation. Therefore, after completing the DFS algorithm, the base station node will allocate resources according to the multipath routing existing in the network topology. The multipath transmission method under the centralized MAC layer protocol is shown in Fig. 2.

In the process of establishing multipath, the network-wide connection matrix is originally maintained in the centralized MAC layer protocol, so no additional overhead is required. At the same time, the network layer no longer needs to establish and maintain routes through RREQ and RREP messages. Therefore, compared with the algorithm for establishing multipath in the network layer, our multipath algorithm greatly reduces the routing overhead.

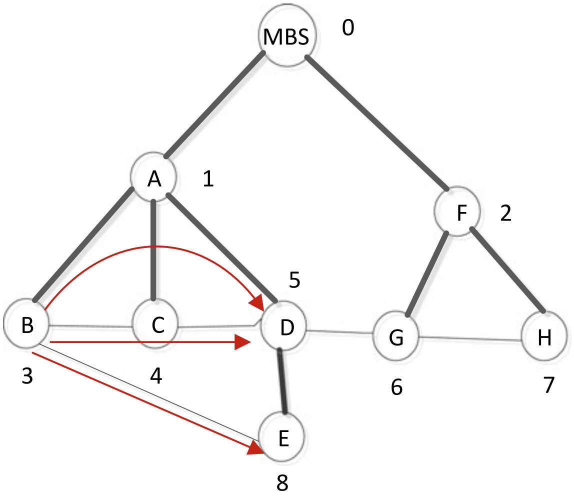


Fig. 2 Multipath transmission based on centralized MAC layer protocol

3.2 Network Coding Transmission Scheme

The coding method in this paper is random linear network coding. Firstly, a batch of data packets put together by the source node for random linear network coding is defined as a generation.

When the source node needs to send data, firstly, it buffers the data packets to be sent in the sending queue. When the number of packets reaches K, the source node will randomly select K encoding coefficients from the finite field F_q, and then encode the K original packets into one encoding packet. Assuming that the total number of encoded packets sent by the source node is $K + n$, the source node needs to continue to select encoding coefficients from the finite field for encoding according to the above encoding operation until $K + n$ encoded packets are generated. $K + n$ encoded packets are transmitted from the source node according to the multipath model shown in Fig. 1, and the relay node receives and forwards the encoded packets to the next-hop node. According to the knowledge of linear algebra, the destination node needs to successfully receive at least K linearly independent coding packets, that is, it can decode to obtain K original data packets. The codec process can be expressed by formula (1).

$$\begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{K+n} \end{bmatrix} = \begin{bmatrix} \alpha_1^1 & \cdots & \alpha_n^1 \\ \vdots & \ddots & \vdots \\ \alpha_1^{K+n} & \cdots & \alpha_n^{K+n} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_K \end{bmatrix}$$

Table 1 Linear vector independent probability of coefficient vector in different finite fields

F_q	Linear independent probability	F_q	Linear independent probability
2^1	0.288788	2^7	0.992126
2^2	0.688538	2^8	0.996078
2^3	0.859406	2^9	0.998043
2^4	0.933595	2^{10}	0.999022
2^5	0.967773	2^{11}	0.999511
2^6	0.984131	2^{12}	0.999756

$$\Rightarrow \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_K \end{bmatrix} = \begin{bmatrix} \alpha_1^1 & \cdots & \alpha_n^1 \\ \vdots & \ddots & \vdots \\ \alpha_1^{K+n} & \cdots & \alpha_n^{K+n} \end{bmatrix}^{-1} \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{K+n} \end{bmatrix} \tag{1}$$

Among them, the choice of the finite field F_q will determine the probability that the coding coefficient is linearly independent, thus affecting the probability of the destination node successfully decoding. If the value of F_q is too large, it will cause a lot of unnecessary coding overhead; if the value of F_q is too small, it will lead to a high correlation of coding coefficients, even if the destination node receives enough coding packets, the probability of successful decoding is still small. Table 1 lists the relationship between the finite field and the linearly independent probability of the coefficient vector. It can be seen that when $F_q = 2^8$, the linearly independent probability of the encoding coefficient reaches 99.6%, and each encoding coefficient occupies only 1 byte [7]. Therefore, we choose $F_q = 2^8$ as the finite field of random linear network coding in this paper.

3.3 Dynamic Load Balancing Strategy

The dynamic load balancing strategy proposed in this paper performs load balancing according to the packet loss rate p_i of different paths. The path packet loss rate p_i is obtained through the feedback mechanism. In addition to the confirmation byte, two bytes need to be added in the ACK message of the feedback mechanism, where one byte records the number of data packets received by the node from the previous-hop node, another byte records the packet loss rate from this node to the next-hop node. The specific process is as follows:

After decoding, the destination node replies with an ACK message to the source node on each path, where the number of data packets is filled in the number of packets received from the current path, and the packet loss rate is filled in null.

After receiving the ACK message, the relay node i calculates the packet loss rate of the link from node i to node $i + 1$ based on the number of data packets recorded in the ACK and the number of data packets received by the node in the previous transmission stage. Then node i fills the calculation result into the packet loss rate field and forwards the ACK message to node $i-1$. After receiving the ACK message, the source node can know the packet loss rate of the path from the ACK message of each path. When the encoded packet is sent in the next stage, load balancing can be performed accordingly.

In the load balancing algorithm, the distribution of data packets on each path is determined according to the packet loss rate of each path:

$$N_i = \frac{1-p_i}{\sum_{i=1}^{k}(1-p_i)} \cdot N \tag{2}$$

where $\sum_{i=1}^{k} N_i = N$, N_i is the number of packets transmitted on path i, and N is the total number of packets to be transmitted.

4 The Performance Model of MPNC

Based on the MPNC transmission scheme designed in Sect. 3, we will analyze the performance and overhead of the MPNC during data transmission in this section. In this paper, the reliability index SDR is defined as the probability that the data sent by the source node is successfully received by the destination node.

4.1 Reliability Model Analysis

In multipath transmission, when the source node needs to send data to the destination node, it establishes multiple paths between the source node and the destination node firstly, and then the source node sends the data packet to the next-hop node. The relay node forwards the received data packets, and finally, the destination node combines the data packets from multiple paths in order to obtain the original data.

We assume that the average link bit error rate for hop i is P_{bi}. Since the reception of each bit of data is independent of each other, the probability that any node receives l-bit data correctly is shown in formula (3):

$$P = (1 - P_{bi})^l, 0 \leq P_{bi} \leq 1 \tag{3}$$

It can be seen from formula (3) that the probability of a coded packet of length l being successfully transmitted at hop i is $P_i = (1 - P_{bi})^l$, so the probability of a data packet being correctly transmitted on the ith path with an average hop count of h can be expressed as:

$$P_{s,i} = \prod_{i=1}^{h} P_i = \prod_{i=1}^{h} (1 - P_{bi})^l \tag{4}$$

When the number of data packets sent by the source node on path i is m, the number of the data packet received by the destination node from this path is denoted by a_i. Since the transmission of each data packet is independent of each other, the number a_i of data packets received from path i follows the binomial distribution with parameter $\left(P_{s,i}, m\right)$, as shown in formula (5):

$$P_{a_i,i} = C_m^{a_i}\left(1 - P_{s,i}\right)^{m-a_i} P_{s,i}^{a_i} \tag{5}$$

When there are k paths to send encoded packets in parallel, use s to represent the sum of all the path data packets received by the destination node, and use the P_s to represent the probability of s to obtain formula (6):

$$P_s = \sum_{a_i=0}^{m}\left(\prod_{i=1}^{k} P_{a_i,i}\right) \tag{6}$$

where the constraints are $\sum_{i=1}^{k} a_i = s$, $\sum_{i=1}^{k} m = K + n.$

Because the encoded packets may be lost during transmission and related to coefficients during decoding, the total number of encoded packets sent by the source node needs to be greater than the number of data packets before encoding. Only then can the destination node decode the original data and ensure the correct transmission of the data. According to the uncorrelation of the coefficient matrix, the relationship between the number of encoded packets received by the destination node and the decoding probability can be solved as shown in formula (7):

$$P_d = \begin{cases} \prod_{i=0}^{K-1}\left(\frac{q^K - q^i}{q^K}\right) = \prod_{i=0}^{K-1}\left(1 - q^{i-s}\right), & s \geq K \\ 0, & s < K \end{cases} \tag{7}$$

Since the finite field selected in this paper is $F_q = 2^8$, it can be seen from Table 1 that the linearly independent probability can reach 99.6%. Therefore, it can be approximated that when the number of encoded packets received by the destination node is K, K original data packets can be solved.

Therefore, when the average number of hops is h, the reliability of data transmission using the MPNC scheme in a multi-hop scenario can be expressed as formula (8):

$$SDR = 1 - \sum_{s=0}^{K-1} p_S \tag{8}$$

That is, the number of original data packets in a generation is K, the number of encoded packets sent by the source node is $K + n$, and the probability of the number of data packets received by the destination node is greater than or equal to K.

4.2 Overhead Model Analysis

The multipath of the MPNC scheme does not depend on the network layer sending RREQ and RREP, but the topology connection matrix maintained by the wireless ad hoc network protocol. Therefore, when calculating the overhead of the MPNC scheme, the cost of establishing and maintaining the route is negligible. The main source of the overhead is the coding cost of network coding.

In the process of using network coding for data transmission, the coding overhead mainly includes two parts: first, the length of the coding coefficient; second, in order to avoid the loss of the encoded packet in transmission and the linear correlation of the decoding coefficient matrix at the destination node, we need to send redundant encoding packets to improve reliability. To measure the size of the overhead by the ratio of the length of the overhead to the total length of the transmitted data, the overhead can be expressed as follows:

$$L = \frac{K\left(K\left(\log_2 q\right)\right)+n\left(l+K\left(\log_2 q\right)\right)}{\left(l+K\left(\log_2 q\right)\right)(K+n)} \tag{9}$$

where K is the number of original data packets in a generation and is also the number of encoded packets that need to be sent at least. n represents the number of encoded packets redundantly transmitted to overcome channel packet loss and decoding failure, q represents the size of the finite field, l represents the length of the original data packet, and $K + n$ represents the total number of encoded packets sent.

5 Simulation Analysis

In this section, we use MATLAB simulation software to evaluate the performance of the proposed MPNC reliable transmission scheme for wireless ad hoc networks. By comparing with the two commonly used reliable transmission scheme (FEC scheme and ARQ scheme), we investigate its performance in terms of both reliability and overhead.

Among them, the length of each data packet information unit is set to 1024 bits, and the NACK message of the ARQ scheme is set to 24 bits according to the size of the packet header. Therefore, $l = 1024$ bits, $l_{NACK} = 24$ bits, the size of each packet of the FEC scheme is 256 bits.

5.1 *Effect of Scheme Parameters on Overhead*

The parameters that affect the overhead of different schemes are different. In order to facilitate the comparison of the overhead of different schemes, we convert the actual overhead size to the overhead ratio for comparison. The overhead ratio is defined as the ratio of the length of redundant data to the total length of the transmitted data. The parameters of each scheme are the number of FEC error correction bits, the maximum number of ARQ retransmission, and the number of MPNC redundant coded packets.

Figure 3 reflects the relationship between the overhead of several reliable transmission schemes and the scheme parameters. As can be seen from the figure, the FEC scheme only adds several redundant bits to the original data to be transmitted to achieve error correction at the receiving end, and the overhead is relatively small. Both the ARQ scheme and the MPNC scheme are redundant to the entire data packet, so the overhead is relatively large. It can be seen from Fig. 3 that the error correction bit of the FEC mechanism is within 10 and the relationship between the overhead and the error correction bit has a linear growth trend. Although the ARQ scheme and the MPNC scheme have small overheads when the scheme parameters are small, it is necessary to analyze whether the reliability at this time can meet the transmission requirements.

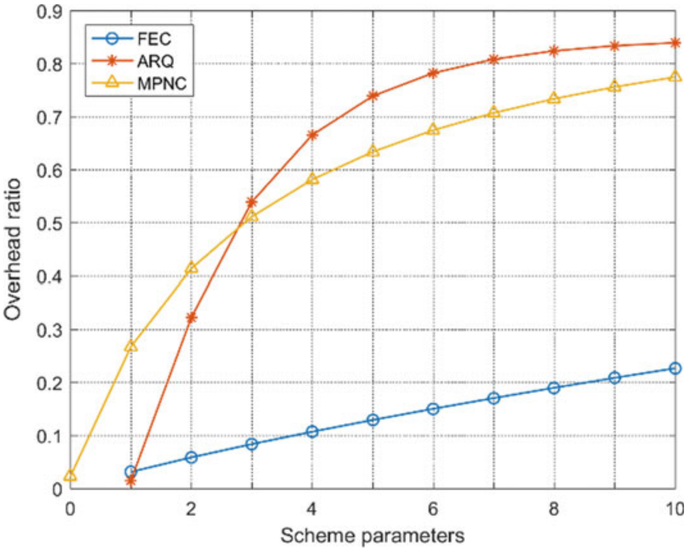


Fig. 3 The relationship between overhead and scheme parameters under different schemes

5.2 Reliability Comparison of Different Scheme in Single-Hop Scenarios

In this section, we first analyze the reliability performance and overhead of several schemes in a single-hop scenario. The number of multipath is set to 3.

Figure 4 is the result of the reliability compared with the multipath transmission strategy when the MPNC scheme has the lowest overhead. It can be seen from the figure that the successful transmission probability of the MPNC scheme and the multipath scheme is decreasing as the link quality decreases. The difference is that the successful probability of multipath transmission drops faster than that of MPNC. When the link bit error rate is 10^{-4}, the reliability of MPNC is basically maintained at 100%, but multipath transmission can only reach 70%.

Figures 5 and 6 are the reliability comparison results of the three schemes when the overhead is 0.1 and the overhead is 0.4. It can be seen from the figure that as the bit error rate increases, the reliability gradually decreases. Among them, the MPNC and the ARQ are a redundant transmission of the entire data packet, so under the condition of fixed overhead, when the bit error rate increases to a certain degree (bit error rate 10^{-4} when the overhead is 0.1, and bit error rate 5×10^{-3} when the overhead is 0.4), the reliability will drop suddenly. This is because the bit error rate at this time cannot meet the requirement of correctly transmitting a packet, and the overhead is fixed, excessive overhead cannot be used. The FEC only adds a small amount of redundancy to the packet, so the lower limit of the overhead is lower and the reliability performance is better.

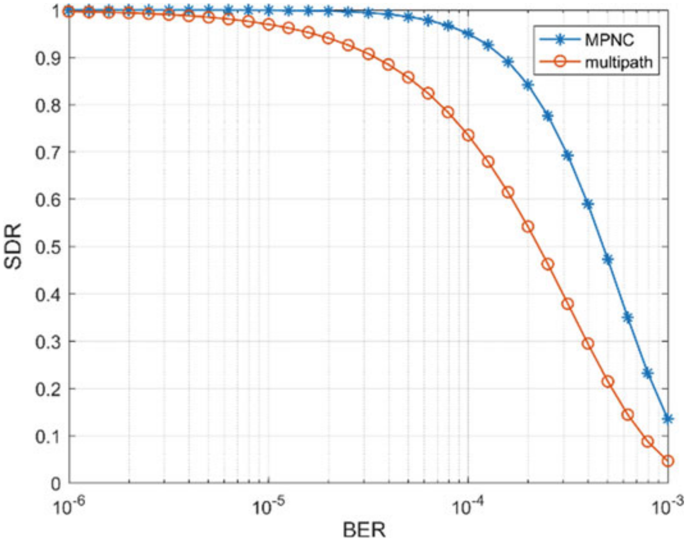


Fig. 4 Reliability comparison under different path models

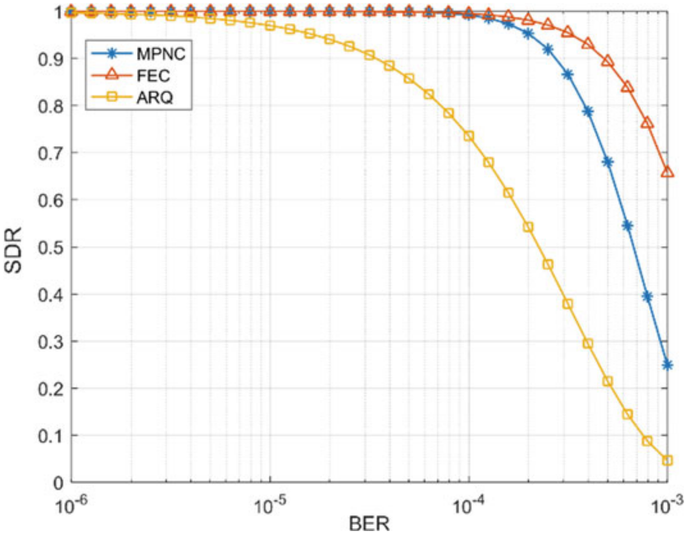


Fig. 5 Reliability comparison of three schemes when the overhead is 0.1

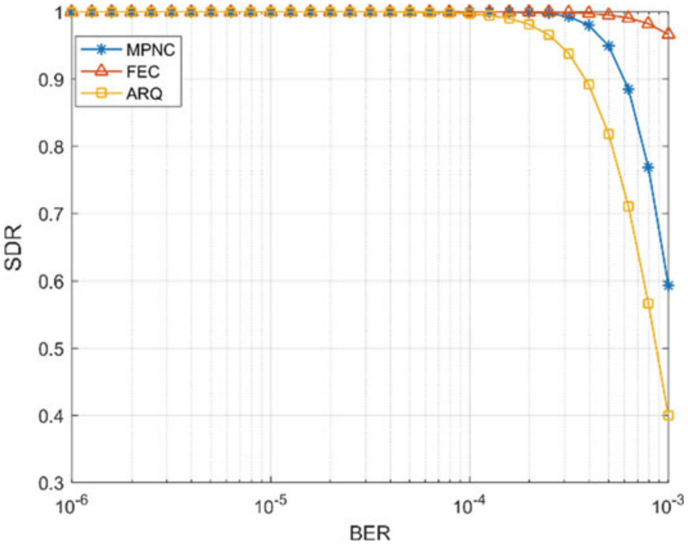


Fig. 6 Reliability comparison of three schemes when the overhead is 0.4

5.3 Reliability Comparison of Different Scheme in Multi-hop Scenarios

In this section, we select three sets of minimum overhead parameters that meet reliability requirements under different channel conditions, and then compare and analyze their performance in multi-hop scenarios, the multi-hop simulation parameter table is shown in Table 2.

Figures 7, 8 and 9 show the results of the reliability multi-hop simulation of the three schemes. It can be seen from the simulation results that in a multi-hop network, as the number of hops increases, the reliability of the three schemes will decrease. Among them, when the bit error rate is low (when the maximum bit error rate is set to 10^{-5} and 10^{-4}), the multi-hop reliability of MPNC performs best, because the correct transmission of packets in MPNC does not depend on the successful transmission of specific packets, but the number of successfully transmitted in a group of packets can meet the decodable requirements. In multi-hop transmission, the packets received by each hop can decode the original data packet within the required range, so the reliability is affected by multi-hop lower. The ARQ also retransmits the entire data

Table 2 Multi-hop simulation parameter table

Maximum bit error rate	FEC	ARQ	MPNC
10^{-3}	4	6	8
10^{-4}	2	3	4
10^{-5}	1	2	2

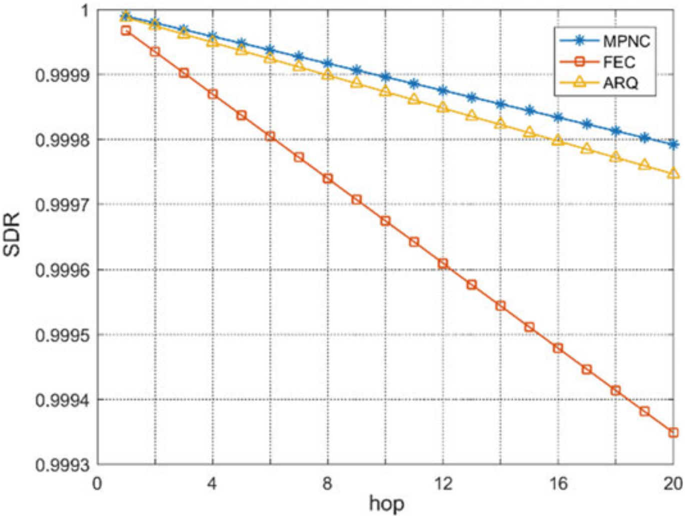


Fig. 7 Reliability changing with hops at max $P_{bi} = 10^{-5}$

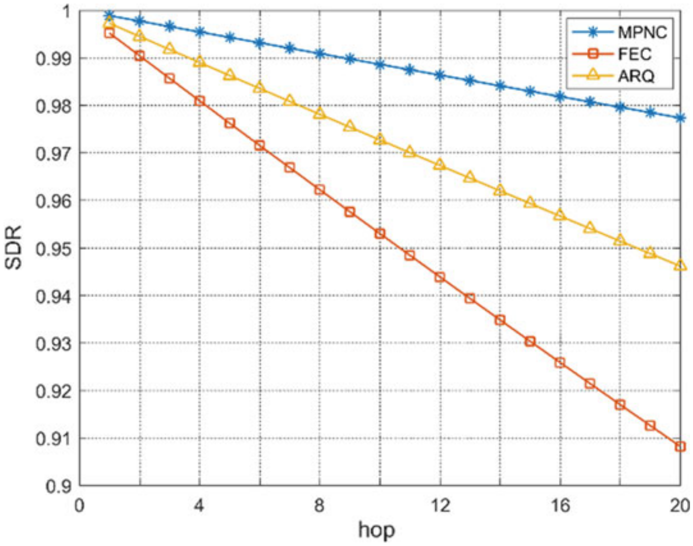


Fig. 8 Reliability changing with hops at max $P_{bi} = 10^{-4}$

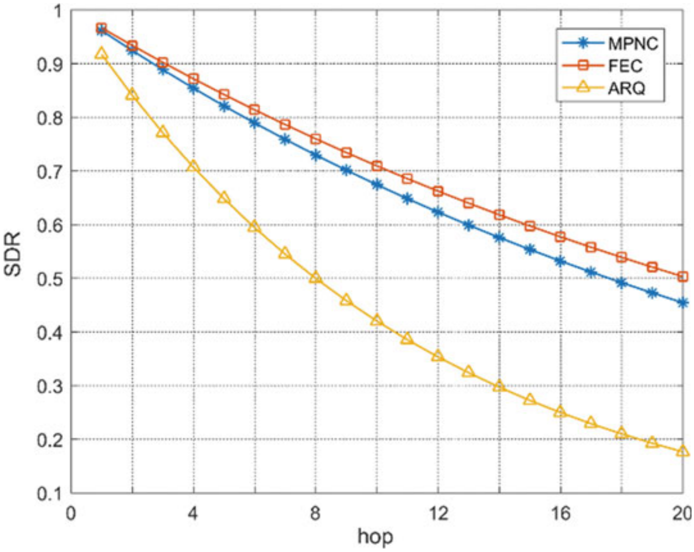


Fig. 9 Reliability changing with hops at max $P_{bi} = 10^{-3}$

packet, when the channel quality is good, the reliability can also be met within the limit of the number of retransmissions. However, when the channel quality is poor, multiple retransmissions of a certain hop will cause the overhead is rising bestially. When the channel quality is good, the error correction bits selected by the FEC can meet the requirements of a single-hop transmission, but the multi-hop transmission is not applicable due to the fluctuation of the link quality, which makes the reliability lower. In addition, when the link quality is poor, ARQ excessively relies on the correct transmission of the entire packet, which leads to a sharp decline in reliability.

6 Conclusion

In this paper, we propose a reliable transmission scheme MPNC for wireless ad hoc networks, which mainly includes the multipath algorithm based on MAC layer protocol, the transmission scheme of network coding, and the design of load balancing strategy. Then, the performance and overhead of the MPNC are modeled and analyzed. Finally, simulations verify the reliability of MPNC in multi-hop networks. The results show that the MPNC scheme can significantly improve the reliability of multi-hop networks over other reliable transmission schemes at fixed overhead. In the future, we will further analyze the applicability of MPNC in actual engineering and implement and verify the engineering.

References

1. K.L. Ang, J.K.P. Seng, Application specific internet of things (ASIoTs): taxonomy, applications, use case and future directions. IEEE Access **7**, 56577–56590 (2019)
2. W. Benrhaiem, A.S. Hafid, P.K. Sahu, Multi-hop reliability for broadcast-based VANET in city environments, in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur (2016), pp. 1–6
3. V. Saritha, P.V. Krishna, I. Alagiri, V.M. Viswanatham, M.S. Obaidat, Efficient multipath routing protocol with quality of service for mobile ad hoc networks, in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO (2018) pp. 1–6
4. S.K. Singh, T. Das, A. Jukan, A survey on internet multipath routing and provisioning. IEEE Commun. Surv. Tutorials **17**(4), 2157–2175 (Fourth quarter 2015)
5. L. Sassatelli, A. Ali, M. Panda, T. Chahed, E. Altman, Reliable transport in delay-tolerant networks with opportunistic routing. IEEE Trans. Wireless Commun. **13**(10), 5546–5557 (2014)
6. R. Mohammadi, A. Ghaffari, Optimizing reliability through network coding in wireless multimedia sensor networks. India Sci. Techn. **8**, 834 (2015)
7. W. Xu, M. Jiang, F. Tang, Y. Yang, Network coding-based multi-path routing algorithm in two-layered satellite networks. IET Commun. **12**(1), 2–8 (2018)
8. X. Liu, M. Médard, W. Li, Network-coding-based multipath transmission in software-defined fiber-Wireless networks, in *2016 IEEE NetSoft Conference and Workshops (NetSoft)*, Seoul (2016), pp. 171–174
9. F. Tang, H. Zhang, L.T. Yang, Multipath cooperative routing with efficient acknowledgement for LEO satellite networks. IEEE Trans. Mob. Comput. **18**(1), 179–192 (2019)

10. Y. Zhang, P. Dong, X. Du, H. Luo, T. Zheng, M. Guizani, BNNC: improving performance of multipath transmission in heterogeneous vehicular networks. IEEE Access **7**, 158113–158125 (2019)
11. Q. Luo, J. Wang, S. Liu, AeroMRP: a multipath reliable transport protocol for aeronautical ad hoc networks. IEEE Internet Things J. **6**(2), 3399–3410 (2019)

Multi-step Time Series Forecasting of Bus Passenger Flow with Deep Learning Methods

Feng Jiao, Lei Huang, and Zetian Gao

Abstract Currently, bus is the major transportation option of the public, with nearly 9 million passengers travelling by bus every day in Beijing, with a result that the bus transportation system in Beijing has experienced huge challenges due to the large volumes of passenger flows. To solve the issues, it is necessary to predict the short-term passenger flow in an accurate way, which allows the schedule system of Beijing Public Transport Corporation to be more efficient, and then to provide better passenger services. In this study, the first step is to clean the bus and weather data and fuse them into a multi-dimensional data set. Then, the bus route 651 was chosen as the research objective, 5 min as time step in prediction. The research built one-step and multi-step prediction models by using LSTM and GRU. In the final step, we would evaluate the prediction performance between distinct prediction models with different hyperparameters. The result reveals that LSTM performs better in multi-step prediction model for route 651.

Keywords Bus passenger flow · Multi-step time series forecasting · LSTM · GRU

1 Introduction

According to the 'Beijing transportation annual report in 2019', public buses are the main choice of vehicles for citizens traveling and commuting and the buses' involvement contributes significantly to the transportation system. There were 24,076 public buses (including electrical bus), 888 routes in Beijing in 2018. The total

F. Jiao (✉) · L. Huang
Department of Information Management School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 15120625@bjtu.edu.cn

L. Huang
e-mail: lhuang@bjtu.edu.cn

Z. Gao
Warwick Manufacturing Group, Warwick University, Coventry, UK
e-mail: Zetian.Gao@warwick.ac.uk

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_38

number of passengers reached 3.19 billion this year, and 9.3 million per day. Thereby, how to control and manage the huge system accurately, effectively and efficiently is a matter of great urgency, which needs to be paid more attention to it. In terms of forecasting passenger flow, it is a very challenging task, because it relies on a plethora of internal and external factors, such as routes, stops, peak and off-peak periods, workdays, weather, special events and so forth. However, the bus company who operates and allocates the bus depends heavily on historical experiences currently. Therefore, in order to improve the efficiency of bus scheduling and management to decrease waiting time of passengers and satisfy their experience, the state of the art technology, Big Data and Deep Learning are utilized, and it is combined with bus data and external data (weather) to explore and analyze the factors and characteristics of passenger flow, and to predict short-term passenger flow more precisely becomes increasingly important.

A multitude of studies were oriented on analyzing passenger flow and predicting the short-term flow. Studies that were focused on predicting the flow of subway are relatively comprehensive, while studies that were oriented on the public bus and electrical bus still need more researches. For instance, Bai et al. [1] used the Deep Belief Network (DBN) to establish a prediction model to tackle the issue of forecasting short-term passenger flow of public buses. Compared to the classic parametric and non-parametric approaches, this model performed better in predicting the passenger flow. Han et al. employed a hybrid optimized LSTM network to predict the actual passenger flow in Qingdao, China and compare the prediction results with those obtained by non-hybrid LSTM models and conventional methods [2]. Toque et al. investigated the short-term multi-step ahead forecasting (t + 1, …, t + 8) of passenger demand in Paris with GRU, aggregated by time step of 15 min [3]. Sun et al. presented a novel adaptive ensemble (AdaEnsemble) learning approach to accurately forecast the volume of metro passenger flow [4].

The main purpose of the study is to use multi-source data sets and then build a model by utilizing Deep Learning to forecast and analyze short-term passenger flow. The main research contents are:

Clean the IC card data set, bus route data set, bus stop data set and weather data set respectively; and then merge them; then calculate each passenger flow with 5 min time step.

Leverage Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to construct a multi-step time series predicting model for short-term public buses passenger flow. This model could capture non-linear characteristics accurately and predict short-term passenger flow between distinct bus routes under the time granularity of 5 min; identify preferable predicting models of short-term passenger flow by comparing each model.

2 Multi-step Time Series Forecasting Model

2.1 Data Source

In this study, a real 3 months IC card data set (March, July, September in 2017), bus route data set, bus stop data set, and weather data set which correspond to same periods, are collected from Beijing Public Transportation Corporation (BPTC) and National Oceanic and Atmospheric Administration (NOAA). Interpretation of each group of data set is listed in Table 1.

1. *IC Card Data Set*
   At the moment, the transportation system in Beijing has leveraged a segment pricing strategy. When passengers take a bus, they need to swipe the cards during their entry and exit respectively. Therefore, a large amount of digital footprints is recorded and generated in their IC cards, and the cards would contain records with getting-on and off (including when and where the passengers get on and off).
2. *Bus Route Data Set*
   The bus route data set is extracted from the network management system of BPTC. The data set depicts the fundamental information of whole bus routes in operation.
3. *Bus Stop Data Set*
   The bus stop data set is also from the network management system of BPTC. The information of the data set contains longitudes and latitudes of bus stops, routes that the stop belongs to, the starts and ends of the routes. In addition, each stop could correspond to its specific route exclusively.
4. *Weather Data Set*
   Weather data set sourced from NOAA in Beijing monitoring points. The data set mainly includes records from March, July and September in 2017. The frequency of collection is 3 h per time. And the main index contains temperature, humidity, wind power, wind direction and so forth.

2.2 Data Preprocessing

In order to a better match between data and the predicting model, it is necessary to preprocess the multivariate data before performing the prediction, which could improve the accuracy of prediction results of short-term passenger flow of public buses. Since the data are collected from different systems and platforms, the collected data might have some inherent issues, such as different data structure and granularity, and partial data loss. To address the issues and improve the accuracy of the predicting model, data preprocessing is a fundamental approach. After that, merge and match the multivariate data together.

Table 1 Data sources

Data sets	Basic information	Data/Records numbers	Data size
IC card data set	Records that passengers swipe IC cards during their entry and exit	649,807,258	64.9 G
Bus route data set	Basic information of bus routes, including route numbers, stop numbers, etc.	1353	352 K
Bus stop data set	Basic information of bus stops, including bus stop numbers, bus stop names, and related routes	54,399	14.1 M
Weather data set	Weather information in corresponding time, including temperature, humidity, wind power etc.	8 records/day	3 KB/day

1. *Data Cleaning*
   IC card data set: In some records, entry and exit are at the same stop. The abnormal data accounts for a small proportion of total data and affects little to the prediction model in reality. Thereby, this kind of data would be purged directly. The field 'Tradetype' contains abnormal trading information, such as blacklist and card lock. In this respect, normal trading information remains exclusively. As for dealing with the basic route and stop information and weather data, the method is to delete irrelevant records.
2. *Data Fusion*
   After data cleaning, all the data sets were joined together by the public fields such as 'XLDM' (the route ID), stop serial number and time. On the basis of this, daily IC card data was ordered by time series at first, then the passenger flow was calculated every 5 min in each route separately. In order to display in a convenient way, the study will select the bus route 651 in March as the research object. As depicted in Table 2, it shows the fusion data of the number of passenger and weather data in the corresponding time, while Table 3 explains the meaning of each variable in Table 2.

Table 2 Example of fusion data

Volume	T	Po	P	U	Ff	VV	RRR
17	−5.5	760.7	767.3	90	1	0.5	0
28	−5.5	760.7	767.3	90	1	0.5	0
62	−5.5	760.7	767.3	90	1	0.5	0
128	−5.5	760.7	767.3	90	1	0.5	0
162	−5.5	760.7	767.3	90	1	0.5	0

Table 3 Field description

Name	Description
Volume	Passenger flow in 5 min
T	Atmospheric temperature at 2 m above ground (Celsius)
Po	Atmospheric pressure at weather station level (mmHg)
P	Average atmospheric pressure at sea level (mmHg)
U	Relative humidity at 2 m above ground (%)
Ff	Average wind speed at 10–12 m above ground within 10 min before observation (meters per second)
VV	Horizontal visibility (km)
RRR	Precipitation (mm)

2.3 Multi-step Time Series Forecasting Model

As is shown in Fig. 1, passenger flow could be regarded as time series where their fluctuation is particularly pertinent to time. Recently, a plethora of models are applied in forecasting short-term passenger flow, such as Neural Networks, ARIMA, Support Vector Machines (SVM), and so forth. In terms of Neural Networks, it could perform extraordinarily in forecasting the number of passengers in the short-term due to its better adaptivity and non-linear characteristics. Deep learning is a more preferable method to deal with large volumes of data and construct models, which is beneficial to better reveal non-linear characteristics and accurately identify the deep regularity in short-term passenger flow. LSTM and GRU are typical time series algorithms in Deep Learning and they could perform better results in short-term prediction. Therefore, the study will use the two models with multi-step time series to predict the passenger flow to identify an appropriate model for forecasting and evaluate the performance of predicting short-term passenger flow by implementing different deep learning models with different network structures.

1. *Overview of LSTM and GRU Models*
   Recurrent Neural Networks (RNNs) could allow the information to remain continuously and constantly. However, the weaknesses of the approach are that it could process a certain amount of short-term dependencies, while vanishing gradient problem and gradient exploding problem might be caused when it deals with long-term dependency problems. Under the circumstance, researchers refined the LSTM on the basis of RNN. Therefore, LSTM not only improves gradient vanishing problem, but also the approach could learn long-term dependency information in time series effectively, and better tackle the delayed issues in time series.
   In a simple RNN, the repeated structural module has only a very simple structure, such as a tanh layer (Fig. 2).

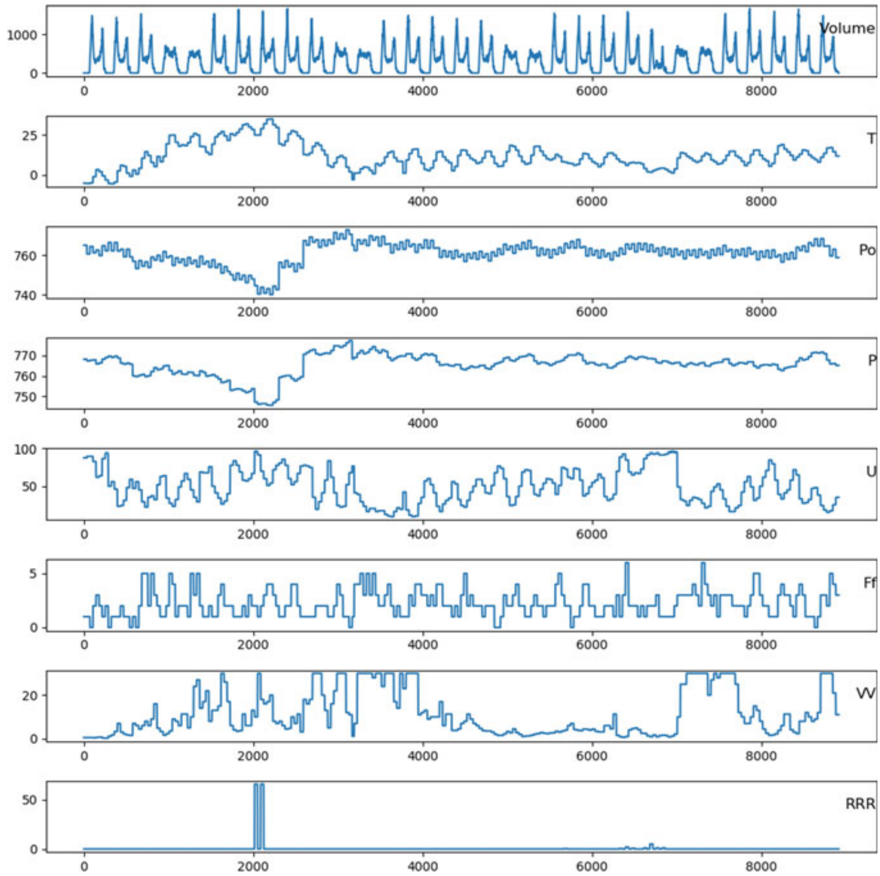


Fig. 1 Visualization of fusion data at 5 min Interval of the route 651 in March

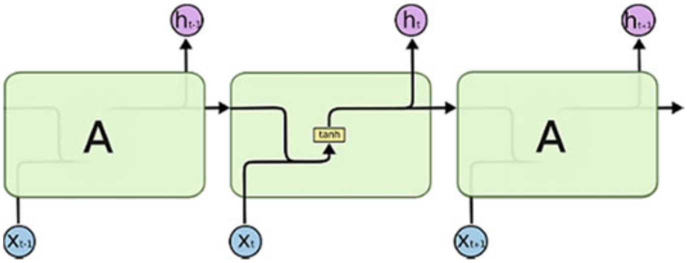


Fig. 2 Illustration of simple RNN

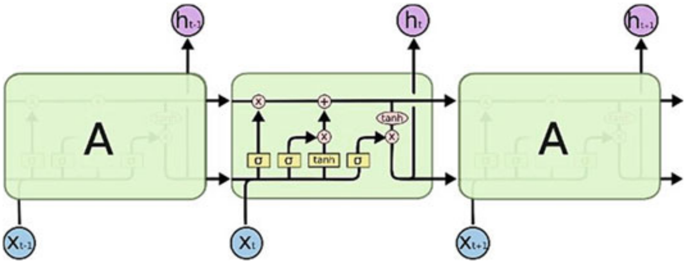


Fig. 3 Illustration of LSTM

On the basis of RNN, LSTM added three gates controlled by the sigmoid unit ([0, 1]) inside each module, they are the Forget Gate, the Input Gate, and the Output Gate (Fig. 3) [5].

The forget gate defines the information that needs to be forgotten by the neuron. It can read the input at the current time and the state h_{t-1} of the hidden layer at the previous time, and finally, output a value between [0, 1] (1 stands for retaining the information, and 0 stands for discarding the information) and assign it to C_{t-1}.

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

The input gate defines the new information stored in the Cell state. It consists of two parts, one of which defines the new candidate value vector through the layer of tanh and the other part defines the value to be input through the layer of sigmoid.

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

And new values will be created to the state of the neuron:

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

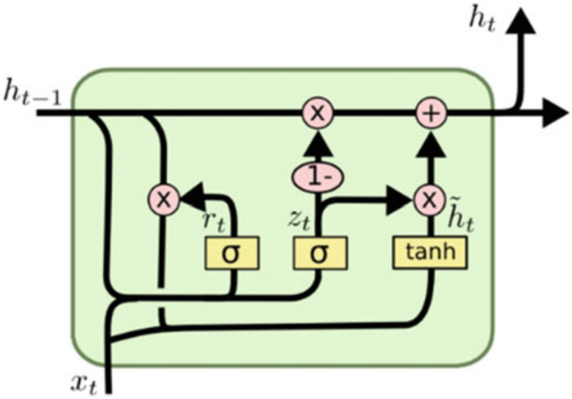
At the same time, the Input Gate will update the Cell state from the original state C_{t-1} to C_t.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4}$$

The output gate defines the values to be output. The output is based on the Cell state. First, use the sigmoid function to determine the information that the Cell state needs to be output. After the Cell state is processed by the tanh layer, it is then multiplied by the output of the sigmoid gate.

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

Fig. 4 Illustration of GRU



$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

where f_t is the forget gate; a i_t is the input gate; o_t is the output gate; W_x is the weight for the respective gate (x) neurons; h_{t-1} is the output of the previous LSTM cell (at time $t-1$); x_t is the input at current time; b_x is the bias for the respective gates (x); C_t is the cell state (memory) at time (t); $\tilde{C}_t$ is the candidate for cell state at time (t).

Based on LSTM, Cho et al. proposed an improved model, Gated Recurrent Unit (GRU) [6]. GRU adds reset gate and update gate on the basis of traditional RNN recurrent neural network, which not only retains the characteristics of RNN "memory" but also enables LSTM to avoid gradient disappearance or explosion. The reset gate is to control the degree of ignoring the information from the previous time. The update gate is to control the extent to which the information at the previous states is brought to the current states.

In Fig. 4, r_t is the reset gate of the GRU network at the current time t; x_t is the input value at the current time t; h_{t-1} is the activation value at the previous time $t-1$; z_t is the update gate of the GRU network at the current time t; h_t is the activation value at the current time t, which is between $\tilde{h}_t$ and h_{t-1}.

2. *Data Processing*

   Before feeding the data into the models, we need to re-frame the time series as supervised time series, which means transforming a sequence to pairs of input and output sequences [7]. We transform time series (Table 2) to three-step supervised series (var1(t), var1(t + 1),var1(t + 2) in Table 4). Then normalize all the features, as shown in Table 5.

3. *Training Set and Testing Set*

   In this study, considering the large data volumes and seasons which will affect the passenger flow, we chose the first 15 days as a training set and the last 15 days as a testing set of each month (March, July, September in 2017).

4. *Multi-step Time Series Short-Time Bus Passenger Flow Forecasting Model with LSTM*

Table 4 Example of re-framed to three-step

var1(t − 1)	var2(t − 1)	var3(t − 1)	var4(t − 1)	var5(t − 1)	var6(t − 1)	var7(t − 1)	var8(t − 1)	var1(t)	var1(t + 1)	var1(t + 2)
1	− 5.5	760.7	767.3	90	4	1	0.5	17	28	62
17	−5.5	760.7	767.3	90	4	1	0.5	28	62	128
28	−5.5	760.7	767.3	90	4	1	0.5	62	128	162
62	−5.5	760.7	767.3	90	4	1	0.5	128	162	228
128	−5.5	760.7	767.3	90	4	1	0.5	162	228	264

Table 5 Normalization

var1(t − 1)	var2(t − 1)	var3(t − 1)	var4(t − 1)	var5(t − 1)	var6(t − 1)	var7(t − 1)	var8(t − 1)	var1(t)	var1(t + 1)	var1(t + 2)
0.010167	0.004926	0.626141	0.685713	0.931034	0.25	0.166667	0.010067	0.016746	0.037081	0.076555
0.016746	0.004926	0.626141	0.685713	0.931034	0.25	0.166667	0.010067	0.037081	0.076555	0.09689
0.037081	0.004926	0.626141	0.685713	0.931034	0.25	0.166667	0.010067	0.076555	0.09689	0.136364
0.076555	0.004926	0.626141	0.685713	0.931034	0.25	0.166667	0.010067	0.09689	0.136364	0.157895
0.09689	0.004926	0.626141	0.685713	0.931034	0.25	0.166667	0.010067	0.136364	0.157895	0.202153

We build this LSTM model with an open-source artificial neural network library called Keras, and take TensorFlow GPU as Keras' backend. Bellows are descriptions of the model's hyperparameter.

(a) *Dropout Layer*
Adding a Dropout layer in the LSTM layer can avoid model overfitting. In the models, the default value of dropout is set to 0.

(b) *Activation Function*
Common activation functions include relu, tanh, sigmoid and etc. In the models, we use tanh as activation function.

(c) *Opitmizer and Loss Function*
In the process of building models, Adam and RMSprop are chosen to be optimizer while MSE and MAS are as Loss function.

(d) *Neurons of LSTM and Training Epochs*
In the models, the number of neurons is set to 50, and the number of epoch (the number of times that all training sets are trained once) is set to 50.

(e) *LR*
LR is the written abbreviation of Learning Rate. Using LR can control the learning progress of the models. In the study, LR is set to 0.001 and 0.005.

5. Multi-step Time Series Short-Time Bus Passenger Flow Forecasting Model with GRU
In order to better compare the performance of LSTM and GRU in short-time passenger flow prediction, Keras + TensorFlow GPU is also used to build the GRU model which has the same hyperparameters and model structure as LSTM. The only difference is that we replace the LSTM layer with the GRU layer.

2.4 Forecast Results and Analysis

In order to clearly evaluate the performance between one-step and multi-step short-term passenger flow forecasting in LSTM and GRU models, we select test set data for 15 days, 5 min as time steps, then compare and visualize the forecasting results of the route 651. In this study, Root Mean Square Error (RMSE) is used as the error evaluation index of the models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

Table 6 reveals the RMSE of prediction results in one-step models with different hyperparameters, while Fig. 5 demonstrates the loss and the minimum RMSE in one-step LSTM models and one-step GRU models respectively.

Table 6 Hyperparameter and RMSE of one-step model

One-step model				
RNN Time series model	Loss function	Optimizer	LR	RMSE
LSTM	MSE	RMSprop	0.001	13.257
LSTM	MAE	RMSprop	0.001	24.024
LSTM	MSE	RMSprop	0.005	32.468
LSTM	MSE	Adam	0.001	14.056
LSTM	MAE	Adam	0.001	20.006
LSTM	MSE	Adam	0.005	11.736
GRU	MSE	RMSprop	0.001	20.557
GRU	MAE	RMSprop	0.001	34.528
GRU	MSE	RMSprop	0.005	29.568
GRU	MSE	Adam	0.001	24.668
GRU	MAE	Adam	0.001	12.002
GRU	MSE	Adam	0.005	15.247

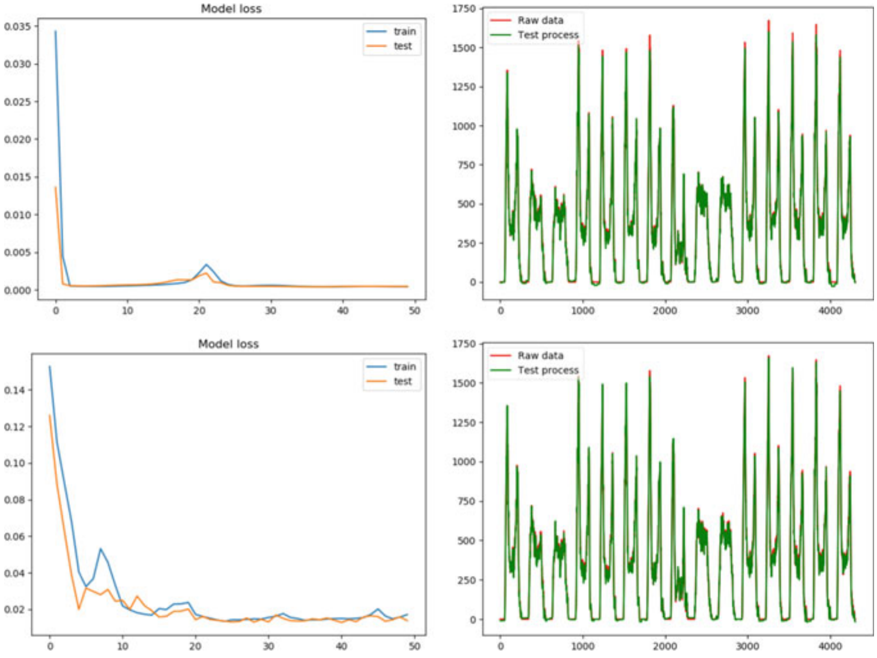


Fig. 5 The comparison of one-step LSTM (RMSE = 11.736) and GRU (RMSE = 12.002)

Table 7 Hyperparameter and RMSE of three-step model

Three-step model				
RNN time series model	Loss function	Optimizr	LR	RMSE
LSTM	MSE	RMSprop	0.001	16.820
LSTM	MAE	RMSprop	0.001	25.037
LSTM	MSE	RMSprop	0.005	39.637
LSTM	MSE	Adam	0.001	20.294
LSTM	MAE	Adam	0.001	10.308
LSTM	MSE	Adam	0.005	34.358
GRU	MSE	RMSprop	0.001	21.086
GRU	MAE	RMSprop	0.001	31.176
GRU	MSE	RMSprop	0.005	50.903
GRU	MSE	Adam	0.001	22.127
GRU	MAE	Adam	0.001	16.823
GRU	MSE	Adam	0.005	27.741

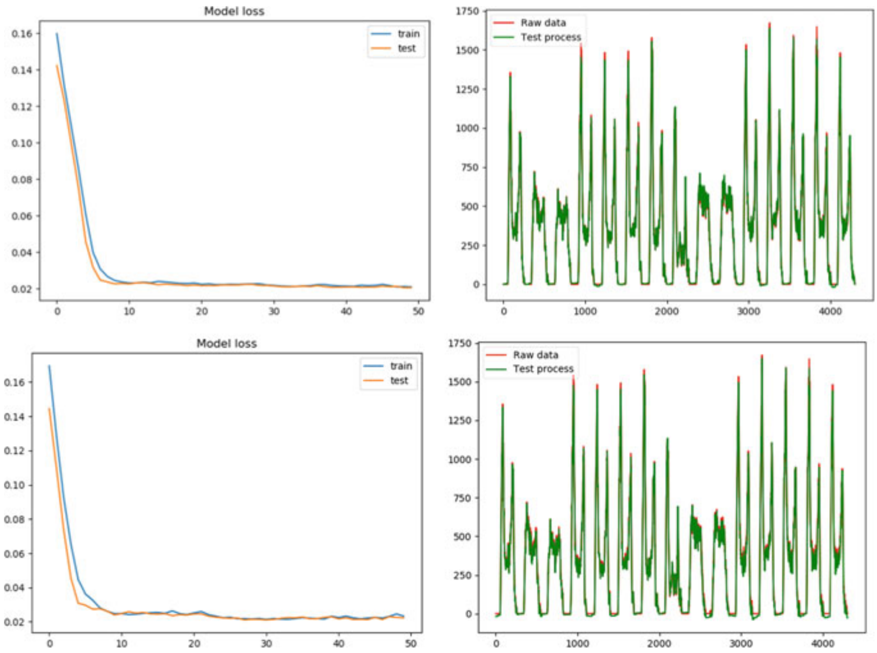


Fig. 6 The comparison of three-step LSTM (RMSE = 10.308) and GRU (RMSE = 16.823)

Table 7 describes the prediction results of the RMSE in three-step models with different hyperparameters, while Fig. 6 demonstrates the loss and the minimum RMSE in three-step LSTM models and three-step GRU models respectively.

As is shown above, in one-step models, LSTM has a smaller prediction error than GRU, but the gap of the error is very narrow. Especially, LSTM has the highest prediction accuracy when loss = MSE, optimizer = Adam, lr = 0.005, while GRU has the highest prediction accuracy when loss = MAE, optimizer = Adam, lr = 0.001. As for three-step models, LSTM also has a smaller prediction error than GRU. Especially, both LSTM and GRU have the highest prediction accuracy when loss = MAE, optimizer = Adam, lr = 0.001.

Besides, the loss in one-step models experiences a sharp fluctuation during the downward and then remains steady. While the loss in multi-step models decreases steadily, which might not overfit under the circumstance. In the research, the epoch is set to 50. If the epoch is set as a viable, the result shows that multi-step models considered the multiple time steps can be more stable, compared to one-step models. Therefore, with regard to predicting the bus passenger flows, multi-step LSTM model can predict the short-term bus passenger flow more accurately.

3 Conclusions

All separate data sets which were collected from distinct systems and platforms were cleaned, the cleaning process includes detecting and correcting the redundant data, missing data and abnormal data. After that, these cleaned data sets were merged into one fusion data set, which is a very fundamental step to build the short-term bus passenger flow prediction model.

We built one-step and three-step short-term passenger flow prediction models based on time series prediction method with deep learning (LSTM and GRU). Then we evaluated the prediction performance of distinct deep learning network models by using different combinations of model hyperparameters.

Based on the prediction results, the model could be helpful to urban transportation planning, passenger transfer, transport schedule, and etc. Besides, the method enables the urban public transportation system to operate effectively and efficiently and allows the trip of citizens to be more convenient and therefore improves their customer service.

Limitation still remains despite some research results are achieved. Further research can be focused on the following aspects:

3.1 Hybrid Short-Term Bus Passenger Flow Prediction Model

Two classic time series prediction models in deep learning were used in the study. The task of further research could leverage combined prediction models to improve the accuracy in short-term bus passenger flow prediction model, since the different prediction models own great functions and characteristics.

3.2 Other Factors Affecting Bus Passenger Flow

In fact, bus passenger flow depends on many factors. That means weather condition is not the only external factor that could be considered. However, travel conditions and special events should also be analyzed when predicting the passenger flow. In further research, we will take these factors into consideration to enable the model to more comprehensive and practical.

References

1. Y. Bai, Z. Sun, B. Zeng, J. Deng, C. Li, A multi-pattern deep fusion model for short-term bus passenger flow forecasting. Appl. Soft Comput. **58**, 669–680 (2017)
2. Y. Han, C. Wang, Y. Ren, S. Wang, H. Zheng, G. Chen, Short-term prediction of bus passenger flow based on a hybrid optimized LSTM network. ISPRS Int. J. Geo-Inf. **8**(9), 366 (2019)
3. F. Toque, E. Come, L. Oukhellou, M. Trepanier, Short-term multi-step ahead forecasting of railway passenger flows during special events with machine learning methods, in *Conference on Advanced Systems in Public Transport and TransitData 2018, CASPT 2018*, Brisbane, Australia, July 2018, p. 15
4. S. Sun, D. Yang, J. Guo, S. Wang, AdaEnsemble Learning Approach for Metro Passenger Flow Forecasting. arXiv.org, 2020. [Online]. Available: https://arxiv.org/abs/2002.07575. Accessed: 10 Apr 2020
5. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. K. Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv.org, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078. Accessed: 10 Apr 2020
7. J. Brownlee, How to convert a time series to a supervised learning problem in python. Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/. Accessed: 10 Apr 2020
8. J. Li, S.X. Pan, L. Huang, X. Zhu, A machine learning based method for customer behavior prediction. Tehnicki Vjesn. Tech. Gaz. **26**(6), 1670–1676 (2019)

Research on Multi-stage Random Distribution Method

Fachao Li, Yueye Zhang, and Chenxia Jin

Abstract With the rapid development of logistics, distribution as an important function, is increasingly concerned. Aiming at the problem of cost increase caused by a large number of delays in actual distribution, this paper proposes multi-stage distribution method based on stochastic programming. Firstly, we propose a multi-stage distribution model in a random environment. Based on the stochastic programming model with the lowest cost, the situation that the whole vehicle distribution requirements are not met in the distribution cycle is discussed. Due to the randomness of demand in distribution, the stochastic programming model will be transformed into the expected value model. Secondly, we simplify the multi-stage distribution model into a node decision-making model, and then use an algorithm to get the decision. Finally, combined with a case, the rationality of the model and the effectiveness of the algorithm are verified.

Keywords Stochastic programming · Delay delivery · Vehicle distribution · Expected value model

1 Introduction

Distribution is the logistics service provided by enterprises when facing the end users directly, which plays a very important role in meeting the needs of users [1]. Article 18 of the 2018 Opinions of the Office of the State Council on Promoting the Coordinated Development of E-commerce and Express Logistics points out: Accelerate

F. Li · Y. Zhang (✉) · C. Jin
School of Economics and Management, Hebei University of Science and Technology,
Shijiazhuang, China
e-mail: 1499132546@qq.com

F. Li
e-mail: lifachao@tsinghua.org.cn

C. Jin
e-mail: jinchenxia2005@126.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_39

the adjustment of the transportation structure, encourage enterprises to comprehensively use information such as e-commerce transactions, logistics and distribution, optimize dispatching, and reduce idle and in-transit time of vehicles. No-load vehicle is a common problem in distribution, which not only increases transportation costs, but also wastes resources so it is not conducive to sustainable development. Therefore, the key to correct decision-making is to reduce the empty load of the vehicle while meeting the customer's needs to the greatest extent, which is also the problem solved in this paper. In the process of enterprise supply chain operation, logistics runs through the whole chain. Distribution is at the end of the whole supply chain and is the final execution link to realize complete supply chain circulation. Among them, the whole vehicle distribution is a kind of distribution method that takes the whole vehicle as the subject matter of logistics service, and makes quick response and on-time distribution according to the customer's order to the requirements of delivery date, delivery place, quality assurance, etc. In the whole vehicle distribution, full-load transportation can make full use of the loading carrier to meet the customer's needs while maximize the benefits, while non-full-load transportation can not make full use of the loading carrier, thus increasing the transportation cost. When loading carriers cannot be fully utilized, logistics enterprises will extend the delivery time to save costs, and extending the delivery time will inevitably cause delayed delivery, thus affecting consumer satisfaction and logistics service quality. Therefore, it is very necessary to use the correct method to make distribution decisions.

In order to reduce the transportation cost and improve the utilization rate of resources, literature [1–3] has done a lot of research on distribution location, continuous and discrete time windows, and port scheduling to reduce distribution costs. Literature [4] optimized the algorithm of the delayed delivery model, considers the relationship between variance, reliability coefficient and reliability degree on the basis of the expected value model, and proposed three tree-based search methods to solve the vehicle scheduling problem. Literature [5] started with two-stage and three-stage delayed delivery, analyzed the characteristics and computational complexity of the model, and gave an N-stage model. Document [6] considers the transportation cost of the third party logistics and optimizes the algorithm. Literature [7] decomposed the production- joint distribution problem into two relatively independent problems of resource optimization and job scheduling, and introduced stochastic programming theory and joint scheduling method to make overall decision. Literature [8] gave the relationship among distribution critical point, cost and compensation for the two-stage problem of delayed distribution, which has been extended to n-stage.

However, there exists randomness in practical problems, which can be solved by random programming. There are three widely accepted stochastic programming methods: expected value model, opportunity constraint model and relevant opportunity model. The above three models are the basis of solving stochastic programming problems today. The common solution is to construct the relevant solution method by integrating random simulation with some intelligent algorithm. Literature [9, 10] designed a solution algorithm for stochastic programming by combining genetic algorithm with stochastic simulation. Literature [11] proposes GA based on priority coding to solve the problem of multiple products and multi-stage supply chain network planning.

Based on the above analysis. To reduce distribution costs, most references focus on scheduling and routing problems, but they all ignore the demand consideration in random environment. For this reason, in view of the randomness of demand in each stage, this paper adopts discrete convolution method to deal with random variables and fuse random information. Based on the two-stage model, a multi-stage distribution model is established and the expression of surplus are given. Finally, according to the characteristics of the multi-stage random distribution model, the multi-stage random distribution model is simplified to a two-stage one.

2 Preliminary Knowledge

2.1 Expected Value Model

Mathematical expectation is a commonly used tool describing the value of random variables. It has good theoretical properties. Through the expected value, the stochastic programming model can be transformed into an ordinary model:

$$\begin{cases} \max\ E(f(x,\xi)), \\ \text{s.t.}\quad E(g_i(x,\xi)) \le 0, \quad i = 1, 2, \ldots, m. \end{cases} \tag{1}$$

Here $x = (x_1, x_2, \ldots, x_n)$ is the decision vector, $E(f(x,\xi))$ is the mathematical expectation of $f(x,\xi), \xi = (\xi_1, \xi_2, \ldots, \xi_n)$ is a given random vector, $f(x,\xi)$ is a target with a random variable, $g_i(x,\xi) \le 0,\ i = 1, 2, \ldots, m$ is a constraint with a random variable in it.

2.2 Discrete Convolution

Discrete convolution formula is a mathematical tool dealing with discrete variables. It discusses the distribution of two independent discrete variables according to certain rules. It can be defined as follows:

Definition 1 Let $x(i), h(i), i, j = 0, 1, 2, \ldots$ be two discrete sequences. Then $y(k) = \sum_{i=0}^{k} x(i)h(k-i)$ is called the convolution of $x(k)$ and $h(k)$.

Proposition 1 *Let X, Y be two independent random variables, and their distribution law are* $P\{X = i\} = x\{i\}$, $P\{Y = j\} = y\{j\}, i, j = 0, 1, 2, \ldots$. *Then the distribution law of Z is equal to X plus Y is*

$$P\{Z = k\} = \sum_{i=0}^{k} x(i)y(k-i) \tag{2}$$

Proposition 2 *Let X, Y, Z be independent random variables, and their distribution law are* $\{X = i\} = x\{i\}$, $P\{Y = j\} = y\{j\}$, $P\{Z = k\} = z\{k\}$, $i, j, k = 0, 1, 2, \ldots$. *Then the distribution law of* $H = X + Y + Z is$

$$P\{H = l\} = \sum_{i=0}^{l} \sum_{j=0}^{l-i} x(i)y(j)z(l-i-j) \tag{3}$$

Proposition 3 *Let* $X_1, X_2, \ldots, X_n$ *be independent random variables, and their distribution law are* $P\{X_t = i_t\} = x_t(i_t)$, $i_t = 0, 1, 2, \ldots, t = 1, 2, \ldots, n$. *Then the distribution law of* $H = \sum_{t=1}^{k} X_t is$

$$P\{H = l\} = \sum_{i_1=0}^{l} \sum_{i_2=0}^{l-i_1} \cdots \sum_{i_{t-2}=0}^{l-i_1-\cdots-i_{t-2}} x_1(i_1)x_2(i_2)\cdots x_t(l-i_1-\cdots-i_{t-1}) \tag{4}$$

2.3 Two-Stage Distribution Model

In the actual distribution problem, customer demand is a random variable, so the distribution problem is a random planning problem. A distribution cycle can be divided into several stages. At the end of each stage, which is regarded as a node. The transportation tasks of each node can be divided into non full load and full load transportation. Full load transportation can make full use of the carrier's capacity. When the order can meet the requirements of full load, it must be delivered on time. Otherwise, you can choose to postpone the delivery, but you need to pay some compensation to the customer.

According to the actual situation, we make the following assumptions: (1) For each delivery cycle, the order must be completed within the delivery cycle. (2) Each cycle is divided into several stages. During each stage, distribution must be completed when the order can meet the full load requirement. When the order does not meet the full load requirement, the intermediate stage can delay the delivery, with the condition of paying a certain compensation fee to the customer. In the last stage of the cycle, distribution must be completed regardless of whether full load requirements are met (Table 1).

$$Z_0(S_1, \xi_2) = S_1\theta + \mathrm{int}\left(\frac{S_1 + \xi_2}{Q}\right)c + \delta\left(\frac{S_1 + \xi_2}{Q}\right)c \tag{5}$$

Table 1 Symbol table

Q	The maximum capacity of a car, $Q \in N^*$
c	Transportation cost of a truck
θ	Delayed compensation to customer for unit merchandise
S	The quantity that does not meet the full load condition in the first stage,which is called residual quantity, $S = 1, 2, \ldots, Q - 1$
ξ_2	The quantity demanded in stage two,which is a random variable, with values ranging from $i = 1, 2, 3, \ldots, m$
$\Pr(\xi = i)$	Represents the probability of the random variable ξ

$$Z_1(S_1, \xi_2) = c + \operatorname{int}\left(\frac{\xi_2}{Q}\right)c + \delta\left(\frac{\xi_2}{Q}\right)c \tag{6}$$

Here (1) $\delta(t)$ is a function determining whether there is a remainder. When t is an integer, $\delta(t) = 0$, otherwise $\delta(t) = 1$; (2) When the order does not meet the delivery conditions for full vehicles, delivery can be delayed, otherwise, delivery cannot be delayed; (3) $Z_0(S_1, \xi 2)$ is the cost function of non-delivery in the first stage; (4) $Z_1(S_1, \xi_2)$ is the cost function when the first stage is delivered.

Since $Z(S_1, \xi_2)$ is a random variable, there is no simple order relation between random variables, so it cannot be solved directly. Therefore, expected value model is a common method to solve stochastic programming problems under certain strategies.

3 The Expected Value Model of Multi-stage Resource Distribution Problem

Based on the two-stage distribution model, a multi-stage distribution model is established. We assume that S_k is the remaining quantity at the end of the stage k, which is a random variable. When $X_k = 1$, $S_k = \{0\}$, When $X_k = 0$, S_k is a random variable with the value of $0 \sim Q - 1$. The model is stated as follows:

$$\begin{cases} \min\ E\big(f_k(S_{k+1,i})\big) = E\Big(\operatorname{int}\Big(\frac{S_{k.i}(1-X_k)+\xi_{k+1}}{Q}\Big)c + \delta\Big(\frac{S_{k.i}(1-X_k)+\xi_{k+1}}{Q}\Big)c\Big)X_{k+1} \\ +E\Big(\Big[\operatorname{int}\Big(\frac{S_{k.i}(1-X_k)+\xi_{k+1}}{Q}\Big)c + S_{k+1,i}\theta\Big] + S_{k+1,i}\theta\Big)(1 - X_{k+1})X_{k+1} + E\big(f_k(S_{k,i})\big), \\ \text{s.t.}\quad S_{k+1,i} = \operatorname{mod}\big[S_{k.i}(1 - X_k) + \xi_{k+1}, Q\big](1 - X_{k+1}), \\ \qquad c \ge 0, \theta \ge 0, i \in \{0, 1\}. \end{cases}$$

Here $f_k(S_k)$ is the cost function at stage k and before stage k.

The residual quantity of each stage is affected by both the residual quantity of the previous stage and the demand quantity of each stage, so it is a big difficulty to find out the distribution of the residual quantity and the joint distribution of the residual quantity and demand quantity of each stage in the model calculation, and because

Table 2 The quantity demanded at stage k + 1

ξ_{k+1}	0	1	2	3	…	n_{k+1}
$\Pr(\xi_{k+1} = j_{k+1})$	$P_{k+1,\ 0}$	$P_{k+1,\ 1}$	$P_{k+1,\ 2}$	$P_{k+1,\ 3}$	$\cdots$	$P_{k+1,\ n}$

the increase of the number of stages, the cost function will grow exponentially. So in order to solve the above two problems, the multi-stage distribution model is regarded as several two-stage models, the two-stage model can be divided into two types: (1) There is no surplus at the end of the first stage, and the distribution of the surplus at the end of the second stage is determined by the demand at the second stage. (2) There is surplus at the end of the first stage, and the distribution of the surplus at the end of the second stage is determined by the joint distribution of the surplus at the first stage and the demand at the second stage.

When $X_k = 1$, $S_k = \{0\}$, so the surplus at stage $k + 1$ is related to the quantity demanded at stage $k + 1$ and the decision value at stage $k + 1$. Suppose that the distribution of ξ_{k+1} is shown in Table 2.

If $X_{k+1} = 1$ is the decision value of stage $k + 1$, then the cost function of stage $k + 1$ is:

$$E\big(Z(S_{k+1,1})\big) = c \sum_{m=0}^{\mathrm{int}\left(\frac{n_{k+1}}{Q}\right)} \sum_{j_{k+1}=m*Q+1}^{n_{k+1}} P_{k+1,j_{k+1}} \tag{7}$$

If $X_{k+1} = 0$ is the decision value of stage $k + 1$, then the cost function of stage $k + 1$ is:

$$E\big(Z(S_{k+1,0})\big) = c \sum_{m=1}^{\mathrm{int}\left(\frac{n_{k+1}}{Q}\right)} \sum_{j_{k+1}=m*Q}^{n_{k+1}} P_{k+1,j} + \theta \sum_{i=1}^{Q-1} \sum_{m=i}^{Q-1} \sum_{j_{k+1}=0}^{\mathrm{int}\left(\frac{n_{k+1}}{Q}\right)} P_{k+1,Q*j_{k+1}+m} \tag{8}$$

When $X_k = 0$, the surplus in stage k is a random variable, and its distribution is shown in Table 3.

The joint distribution law of S_k and ξ_{k+1} is $T = S_k + \xi_{k+1}$. The probability of random variable T is expressed as P_T:

$$P(S_k + \xi_{k+1} = T) = \sum_{j_{S_k}=0}^{Q-1} \sum_{j_{k+1}=0}^{n_{k+1}} q_{k,j_{S_k}} * P_{k+1,j_{k+1}} \tag{9}$$

Table 3 The value of the remainder at stage *k*

S_k	0	1	2	3	…	$Q-1$
$\Pr(S_k = j_{S_k})$	$q_{k,\ 0}$	$q_{k,\ 1}$	$q_{k,\ 2}$	$q_{k,\ 3}$	…	$q_{k,\ Q-1}$

If $X_{k+1} = 1$ is the decision value of stage $k + 1$, then the cost function of stage $k + 1$ is:

$$E\big(Z(S_{k+1,1})\big) = c \sum_{m=0}^{\mathrm{int}\left(\frac{Q-1+n_{k+1}}{Q}\right)} \sum_{T=m*Q+1}^{Q-1+n_{k+1}} P_T \tag{10}$$

If $X_{k+1} = 0$ is the decision value of stage $k + 1$, then the cost function of stage $k + 1$ is:

$$E\big(Z(S_{k+1,0})\big) = c \sum_{m=1}^{\mathrm{int}\left(\frac{Q-1+n_{k+1}}{Q}\right)} \sum_{T=m*Q}^{Q-1+n_{k+1}} P_T + \theta \sum_{i=1}^{Q-1} \sum_{m=i}^{Q-1} \sum_{T=0}^{\mathrm{int}\left(\frac{Q-1+n_{k+1}}{Q}\right)} P_{Q*T+m} \tag{11}$$

4 Node Residual Quantity Judgment Method

The above-mentioned multi-stage distribution model is established on the basis that the demand is a random variable, and the solution scheme is given. The decision is made according to the cost value in different situations, and the decision value corresponding to the minimum cost is the optimal decision. In this section, the multi-stage distribution model is simplified into a node judgment model, and decisions are made according to the surplus of each stage.

The end of each stage is called a node, and the nodes are divided into intermediate nodes and end nodes. For example, in the three-stage model, the nodes at the end of the first stage and the end of the second stage are called intermediate nodes, and the node at the end of the third stage are called end node. According to the principle of "full car must be sent, non-full can choose not to send", there are two decision methods for the intermediate node, send or not send. There is only one decision principle at the end node, that is, "whatever is needed must be sent". If the future demand of each node in the multi-stage distribution problem is regarded as the whole, then each node can be regarded as the end node, so the multi-stage distribution problem can be simplified into a two-stage distribution problem, then according to the surplus of each node to judge whether the current stage of distribution. As follows:

$$Z_0(S_1, \chi_1) = S_1\theta + \mathrm{int}\left(\frac{S_1 + \chi_1}{Q}\right)c + \delta\left(\frac{S_1 + \chi_1}{Q}\right)c \tag{12}$$

$$Z_1(S_1, \chi_1) = c + \mathrm{int}\left(\frac{\chi_1}{Q}\right)c + \delta\left(\frac{\chi_1}{Q}\right)c \tag{13}$$

$$\begin{cases} \min \; Z = x_1 c + (1-x_1)S_1\theta + \mathrm{int}\left(\dfrac{\chi_1 + (1-x_1)S_1}{Q}\right)c + \delta\left(\dfrac{\chi_1 + (1-x_1)S_1}{Q}\right)c, \\ \text{s.t.} \quad c \geq 0, \theta \geq 0, x_1 = 0, 1. \end{cases}$$

Here $\chi_1 = \sum_{i=2}^{n} \xi_i$ is the sum of all the quantities demanded after the first stage. We use expectation value to represent a random variable, then the distribution cost can be converted to $E(Z_0(S_1, \chi_1))$ and $E(Z_1(S_1, \chi_1))$.

4.1 Resolution Steps

From the model, we can see that the distribution cost of each stage is related to the surplus of the previous stage, and the demand at the end of each stage is the sum of all the demand after the stage, so the solution of multi-stage distribution model can be seen as the cycle of multiple two-stage models. We can use the following algorithm to complete the calculation of n-stage distribution problem.

Step 1: Input the initial parameters.

(a) Given parameters c, θ, Q and the value of demand ξ_i at each stage.
(b) The distribution law of surplus S_1 is obtained by demand ξ_1, and get the distribution of the total demand $\chi_1 = \sum_{i=2}^{3} \xi_i$ through the demand after the first stage.

Step 2: Cost function comparison.

(c) The cost $E(Z_1(S_1, \chi_1))$ of the first stage residual distribution and the cost $E(Z_0(S_1, \chi_1))$ of the first stage residual delayed distribution are calculated.
(d) Cost comparison, if $E(Z_1) > E(Z_0)$, then $x_1 = 0$ and $\chi_2 = \xi_3$, get S_2 by $H_1 = S_1 + \xi_2$. Otherwise, $x_1 = 1$, $\chi_2 = \xi_3$, and get S_2 through ξ_2.
(e) Repeat steps (a) and (b) until the end.

Step 3: Calculate all possible values.
Step 4: Obtain the solution.

For example, suppose a distribution cycle is divided into three stages, the maximum loading capacity of each truck is Q, the single transportation cost of each truck is c, the compensation value for each delayed car being θ, and the demand for each stage is random variables with distribution $P\{\xi_t = i_t\} = p_t(i_t), t = 1, 2, \ldots n_i, i_t = 0, 1, 2 \ldots$. The distribution of the remaining demand in the first stage is $P(S_1 = i_{S_1}) = \sum_{i=0}^{\mathrm{int}(n_1/Q)} p_{1,Q*i+Q-1}$, the total demand in the later stage is the combined demand in the second and third stages, and the distribution is $P(\chi_1 = i_{\chi_1}) = \sum_{i_2=0}^{n_2} \sum_{i_3=0}^{n_3} p_{i_2} * p_{i_3}$, thus we can obtain the first stage model:

$$\begin{cases} \min\ E(Z) = x_1 c + (1 - x_1)E(S_1)\theta + E\left(\begin{bmatrix} \operatorname{int}\left(\dfrac{\chi_1 + (1 - x_1)S_1}{Q}\right)c \\ +\delta\left(\dfrac{\chi_1 + (1 - x_1)S_1}{Q}\right)c \end{bmatrix}\right), \\ \text{s.t.}\quad c \geq 0, \theta \geq 0, x_1 = 0, 1. \end{cases}$$

Find the optimal solution of the first stage. If the optimal solution is $x_1 = 1$, the residual distribution of the second stage is $P(S_2 = i_{S_2}) = \sum_{i_2=0}^{\text{int}(n_2/Q)} p_{2,Q*i_2+Q-1}$. The total demand in the latter stage is the demand in the third stage, and the distribution is $P\{\xi_3 = i_t\} = p_3(i_3), i_3$=1, 2, ... n_3. If the optimal solution is $x_1 = 0$, the residual distribution of the second stage is $P(S_2 = i_{S_2}) = \sum_{i_2=0}^{n_2} \sum_{i_{S_1}=0}^{Q-1} p_{i_2} * p_{i_{S_1}}$. The total demand in the latter stage is the demand in the third stage, and the distribution is $P\{\xi_3 = i_t\} = p_3(i_3), i_3 = 1, 2, \ldots n_3$, obtaining the second stage model:

$$\begin{cases} \min E(Z) = x_2 c + (1 - x_2)E(S_2)\theta + E\left(\begin{bmatrix} \operatorname{int}\left(\frac{\chi_2+(1-x_2)S_2}{Q}\right)c \\ +\delta\left(\frac{\chi_2+(1-x_2)S_2}{Q}\right)c \end{bmatrix}\right) \\ \text{s.t.}\quad c \geq 0, \theta \geq 0, x_2 = 0, 1. \end{cases}$$

The optimal solution of the second stage model is the decision value of the second stage. The distribution cost under each decision can be calculated by taking into formulas (12) and (13). In the third stage, distribution will be carried out regardless of whether it is full or not. If the decision value in the second stage is $x_2 = 1$, the distribution cost is $E(Z_1) = c + E(\text{int}(\xi_3/Q)c + \delta(\xi_3/Q)c)$. If the decision value in the second stage is $x_2 = 0$, the cost is $E(Z_1) = E(S_2)\theta + E(\text{int}((S_2 + \xi_3)/Q)c + \delta((S_2 + \xi_3)/Q)c)$.

5 Case Analysis

This part combines a multi-stage distribution problem with stochastic demand to analyze how to make decisions in each stage.

A car distribution center will deliver a batch of cars in four stages within one month. The random distribution of demand in each stage is shown in Table 4. The total vehicle load of distribution vehicles is 6, and the transportation cost of the whole vehicle is 18,000 yuan. Orders must be completed within the distribution cycle. For each stage of the order, full load must be allocated, not full load can delay the delivery, delay the delivery of each car to pay 2500 yuan of compensation costs, in the face of random demand in each stage, how should the decision makers make decisions in each stage.

According to the judgment method of node residual quantity, the distribution of residual quantity and residual demand at the end of each stage can be obtained. The

Table 4 Distribution of demand at each stage

ξ_i	2	4	6	8	9	10	12
$\Pr(\xi_1 = i_1)$	0.2	0.2	0.1	0.1	0.2	0.1	0.1
$\Pr(\xi_2 = i_2)$	0.2	0.1	0.1	0.2	0.1	0.1	0.2
$\Pr(\xi_3 = i_3)$	0.1	0.15	0.15	0.2	0.15	0.1	0.15
$\Pr(\xi_4 = i_4)$	0.1	0.1	0.1	0.1	0.2	0.1	0.3

Table 5 The remainder of the stage 1

S_1	0	1	2	3	4	5
$\Pr(S_1 = i_{S_1})$	0.2	0	0.3	0.2	0.3	0

first stage is shown in Tables 5 and 6, the second stage is shown in Tables 7 and 8, and the third stage is shown in Table 9. The results are put into formula (12) and (13), and the decision is made after comparing the results, as shown in Table 10.

Table 6 Prediction of demand after the stage 1

χ_1	6	8	10	12	13	14
$\Pr(\chi_1 = i_{\chi_1})$	0.002	0.006	0.0115	0.0205	0.008	0.03
χ_1	15	16	17	18	19	20
$\Pr(\chi_1 = i_{\chi_1})$	0.015	0.0465	0.021	0.0595	0.033	0.0765
χ_1	21	22	23	24	25	26
$\Pr(\chi_1 = i_{\chi_1})$	0.0365	0.084	0.055	0.0795	0.046	0.083
χ_1	27	28	29	30	31	32
$\Pr(\chi_1 = i_{\chi_1})$	0.041	0.0625	0.038	0.0535	0.019	0.0305
χ_1	33	34	36			
$\Pr(\chi_1 = i_{\chi_1})$	0.0195	0.0135	0.009			

Table 7 The remainder of the stage 2

S_2	0	1	2	3	4	5
$\Pr(S_2 = i_{S_2})$	0.3	0	0.4	0.1	0.2	0

Table 8 Prediction of demand after the stage 2

χ_2	4	6	8	10	11	12
$\Pr(\chi_2 = i_{\chi_2})$	0.01	0.025	0.04	0.06	0.035	0.07
χ_2	13	14	15	16	17	18
$\Pr(\chi_2 = i_{\chi_2})$	0.045	0.105	0.045	0.105	0.055	0.12
χ_2	19	20	21	22	24	
$\Pr(\chi_2 = i_{\chi_2})$	0.035	0.085	0.075	0.045	0.045	

Table 9 The remainder of the stage 3

S_3	0	1	2	3	4	5
$\Pr(S_3 = i_{S_3})$	0.265	0.055	0.26	0.075	0.255	0.09

Table 10 A comparison of the results

Decision table	$Z_0(S, \chi)$	$Z_1(S, \chi)$	X
Stage 1	94,410	95,844	1
Stage 2	72,180	69,710	0
Stage 3	48,600	50,320	1

From Table 10, it can be seen that the first and third phases should be delivered in time, the second phase should be delayed, so as to minimize the cost of delivery.

6 Conclusion

Aiming at the problem of multi-stage random allocation, firstly, this paper analyzes the characteristics and shortcomings of the existing random allocation methods. Secondly, proposing a multi-stage allocation model based on the existing two-stage model, and uses discrete convolution formula to process the random information of each stage. Finally, the model is applied to a specific case for analysis. The results show that the discrete convolution formula can effectively deal with random variables in multiple stages, and has the advantages of simple operation, strong interpretability and low computational complexity. At the same time, the multi-stage distribution model can not only obtain the decision of each stage, but also estimate the overall cost. The data is clear and intuitive, which can effectively avoid the influence of subjective consciousness on decision-making. Therefore, the discussion in this paper has laid a foundation for the further establishment of distribution problems in complex environments.

Acknowledgements This work is supported by the National Natural Science Foundation of China (71771078, 71371064), Hebei graduate innovation funding project.

References

1. S. Ropke, D. Pisinger, A unified heuristic for a large class of vehicle routing problems with back hauls. Eur. J. Oper. Res. **171**(3), 750–775 (2006)
2. G. Jeon, H.R. Leep, J.Y. Shim, A vehicle routing problem solved by using a hybrid genetic algorithm. Comput. Ind. Eng. **53**(4), 680–692 (2007)
3. R. Dondo, A. Méndez Carlos, J. Cerdá, Optimal management of logistic activities in multisite environments. Comput. Chem. Eng. **32**(11), 2547–2569 (2008)
4. W. Yuan, J. Xinjia, L.H. Loo, Tree based searching approaches for integrated vehicle dispatching and container allocation in a transshipment hub. Expert Syst. Appl. **15**(74), 139–150 (2017)
5. F.C. Li, C.X. Jin, L. Wang, Applied mathematical modelling quasi-linear stochastic programming model based on expectation and variance and its application in transportation problem. Appl. Math. Model. **7**(38), 1919–1928 (2014)
6. M.N. Amir, M. Mahdavi, Coordinating order acceptance and integrated production-distribution scheduling with batch delivery considering third party logistics distribution. J. Manuf. Syst. **1**(46), 29–45 (2018)
7. M. Du, Two stage optimization method of production distribution joint decision under vehicle restriction. Ph.D. dissertation, Dalian University of Technology, Dalian (2017)
8. L. Zhou, F.C. Li, The two-stage delay distribution method based on compensation mechanism under random environment. Int. J. Internet Manuf. Serv. **6**(1), 19–31 (2019)
9. B.D. Liu, Dependent-chance goal programming and its genetic algorithm based approach. Math. Comput. Model. **24**(7), 43–52 (1996)
10. K. Iwamura, B. Liu, A genetic algorithm for chance constrained programming. J. Inf. Optim. Sci. **17**(2), 40–47 (1996)
11. F. Altiparmak, M. Gen, L. Lin, I. Karaoglan, A steady-state genetic algorithm for multi-product supply chain network design. Comput. Ind. Eng. **56**(2), 521–537 (2007)

Vessel Scheduling and Refueling in Liner Shipping: Modeling Transport of Perishable Assets with a Collaborative Agreement

De-chang Li, Hua-long Yang, and Wei Song

Abstract Attracting new customers and consolidation to form stronger alliances are the widely used strategies by liner shipping companies, which have made some positive impact. Liner shipping companies may incur substantial customer defection without a proper design of vessel schedules for the transportation of perishable assets. This paper proposes a novel mixed integer non-linear mathematical model for the vessel scheduling and refueling strategy problem about transportation of perishable assets with an efficient collaborative agreement in a liner shipping route, which minimizes the total route service cost and clearly simulates the decay of perishable assets on board. The original mixed integer nonlinear model is linearized by a set of piecewise linear secant approximations, and CPLEX is used to solve the linearized mathematical model. A number of numerical experiments are conducted for the AEU6 liner shipping route, served by the CHINA COSCO SHIPPING GROUP liner shipping company. Some insights from numerical experiments are provided in the end.

Keywords Liner shipping schedules · Collaborative agreement · Perishable assets · Optimization model · Refueling strategy

1 Introduction

The containerized cargos are sensitive to many factors from operational and environmental such as: transportation time, temperature, humidity, barometric pressure, and air composition. The perishable assets are wasted due to temperature variations

D. Li · H. Yang (✉) · W. Song
Transportation Engineering College, Dalian Maritime University, Dalian, Liaoning, China
e-mail: hlyang@dlmu.edu.cn

D. Li
e-mail: lidechangsdut@163.com

W. Song
e-mail: songwei_198@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_40

and other factors that may facilitate asset deterioration account for 25% of the total perishable assets every year [1].

Unlike tramp shipping, liner shipping services are similar to public transport services [2], which have four fixed characteristics: fixed routes, fixed ports, fixed freight rates, and fixed sailing schedules, i.e., arrival and departure times at each port of call [3–5]. Schedule design for a liner shipping services is a tactical-level planning decision that is made every three to six months. Schedule is announced in advance to collect Cargos. A port needs to provide services for a lot of liner shipping companies and ships, which can't guarantee the services will be available whenever a ship arrives. The service availability of the ports is the most important factor to be considered in design of vessel schedules. In addition, Global container port-handling and trade tensions are getting more and more complicated. Consequently, the liner shipping company must cooperate with marine container terminal operators in order to obtain more available time windows and handling rates provided by marine container terminal operators.

Company competitiveness and potential monetary losses also influence the design of schedule [6]. A perishable asset should be delivered to the destination before it is of an acceptable quality for the customer [7]. In order to attract potential customers, liner shipping companies have to make necessary alterations in vessel schedules to reduce waste of perishable assets. A higher sailing speed and handling rate will reduce the round-trip journey time and handling time, thereby the number of vessels required, inventory cost and perishable assets cost. However, a higher speed implies a higher bunker cost and handling cost [8]. Consequently, a liner shipping company must balance the trade-off between ship cost, bunker cost, port handling cost, delay penalty cost, inventory cost and perishable assets cost in schedule design, subject to the constraints of vessel service time windows (TW) and handling rates negotiated by the liner shipping company and marine container terminal operators.

In recent decades, most of the studies on liner shipping mostly focused on a tactical level vessel scheduling problem [9]. The vessel scheduling problem aims to optimize the sailing speed at each voyage leg, arrival times, handling times and departure times [10]. However, the studies on vessel scheduling give less attention to bunkering optimization. The bunker cost account for about three quarters of the operating costs [11]. Since fuel price directly determines the speed optimization of the liner shipping, the fuel price and the vessel schedule usually influence each other.

The contributions of the paper are three-folds: first, this study extends the work, conducted by Dulebenets and Ozguven [12], and proposes a novel mixed integer non-linear mathematical model for the vessel scheduling and refueling strategy problem with an efficient collaborative agreement which had been used in the published to date vessel scheduling literature [13, 14]. Second, the choice of bunkering ports and bunkering amount will affect the time either sailing time or vessel handling time. Given the prevalence of the change in quality of asset type in marine transportation due to sailing time and handling time. This is a natural extension to operations research. Due to the addition of perishable assets, the vessel scheduling and refueling strategy would be reprogrammed. Nonetheless, there is little literature considering a chain reaction in these problems. Third, a detailed comparative analysis is conducted to assess advantages of the proposed transporting of perishable assets with an efficient collaborative agreement over the other Existing research and reveal some important managerial insights.

2 Problem Description

This section provides a detailed description of the problem covering areas including: (a) general description; (b) collaborative agreement description; (c) the port handling cost; (d) the late arrival penalty; (e) container inventory cost; and (f) transport of perishable assets; (g) refill ports and bunker cost.

2.1 General Description

If the liner shipping companies according to the weekly ship frequency, a round trip time of the vessel should be an integer times of the week, and the integer should also be the number of vessels configured on the route. When transporting perishable goods, the corruption loss cost of perishable assets is usually positively correlated with the vessel's sailing time and port handling time between the port of origin and port (OD flow), that is, the longer the time, the higher the corruption loss cost. At the same time, it is not difficult to see that the liner transport operating cost and container inventory cost are also positively correlated with the ship's sailing time and port handling time. Since the bunk cost has a cubic relationship with the ship's speed, the bunk cost has a negative relationship with the sailing time. When a high port handling rate is chosen, the handling cost will be higher, but the handling time will be reduced. Therefore, there is a negative correlation between the handling cost and the handling time. In addition, the liner shipping company and each affiliated port of the ship have a time window agreement for the ship arrive/departure the port. If the ship delays in arriving at the port, there will be a delay penalty cost. Obviously, the cost is also positively correlated with the vessel's sailing time and port handling time. The choice of different refueling strategies directly affects the fuel price of each section, thus causing the adjustment of sailing speed. The adjustment of sailing speed affects the arrival time of ships, thus affecting the selection of handling rates and time in port. Then it has a linkage influence on the whole shipping schedule design. Because the corruption of perishable assets is directly related to the time at sea and the time at port, it makes the design of the whole vessel schedule and the selection of handling rates more complicated.

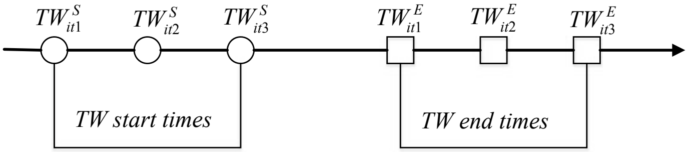


Fig. 1 3 start times and 3 end times scenario at a port

2.2 Collaborative Agreement

We denote by $I = \{1, 2, \ldots, n\}$ the set of ports where vessels have to be visited. A vessel sails between two consecutive ports i and $i + 1$ along voyage leg i. The liner shipping company is able to negotiate a specific agreement with the marine container terminal operators, which has three main contents: (1) a set of vessel arrival TWs $T_i = \{1, 2, \ldots, m_i\}$, $i \in I$ is provided to the liner shipping company at each ports of call; (2) a set of start times $S_{it} = \{1, 2, \ldots, g_{it}\}$, $i \in I$, $t \in T_i$ and end times $E_{it} = \{1, 2, \ldots, o_{it}\}$, $i \in I$, $t \in T_i$ are included in each TW; and (3) a set of handling rates $H_{it} = \{1, 2, \ldots, w_{it}\}$, $i \in I$, $t \in T_i$ is provided to the liner shipping company during each TW.

The arrival TW at each port is divided into two parts: TW_{its}^{S}—is start time s for TW t at port i; and TW_{ite}^{E}—is end time e for TW t at port i, during which a vessel should arrive at port of call. In the example, Fig. 1 demonstrates a scenario, where the marine container terminal operator of port i provides 3 start times and 3 end times for each TW t.

2.3 Port Handling Cost

It is believed that higher handling productivities means higher vessel handling cost. The total port handling cost is expected to pay can be computed as:

$$TPC = \sum_{i \in I} \sum_{t \in T_i} \sum_{h \in H_{it}} C_{ith}^{hc} y_{ith} D_i,$$

where *TPC* is the total port handling cost (USD); y_{ith} is a binary variable used to determine the vessel handling rate.

2.4 Late Arrival Penalty

Vessels, arriving after the end of the selected time of arrival TW may cause service delays for other vessels. The total vessel penalty for late vessel arrivals can be estimated using the following equation:

$$TLP = \sum_{i \in I} C_i^{lc} t_i^l ,$$

where *TLP* is the total late arrival penalty (USD); C_i^{lc} is the unit late vessel arrival penalty at port i (USD/h); and t_i^l is the vessel late arrival hours at port i (h).

2.5 Container Inventory Cost

We are only concerned about the inventory cost at sea was affected by the total vessel sailing time and the total number of containers transported, the value of which can be calculated as [15]:

$$TIC = C^{ic} \sum_{i \in I} Q_i t_i ,$$

where *TIC* is the total container inventory cost (USD); C^{ic} is the unit container inventory cost (USD per TEU per hour); Q_i is the volume of containers transported by vessels at voyage leg i (TEUs); t_i is the sailing time at voyage leg i (h).

2.6 Transport of Perishable Assets

It is assumed that a set of perishable assets $K = \{1, \ldots, r\}$ are transported by the liner shipping company. We denote by $P_{ijk} = \left\{1, \ldots, n_{ijk}\right\}$ the set of ports the vessel is attached in the process of transporting the each asset type k that has an origin port i and a destination port j. The change in quality of asset type k after transportation from the origin port to the destination port can be calculated using the following equation [16]:

$$h\left(t_{ijk}\right) = 1 - e^{-\delta_k t_{ijk}} \quad \forall i, j \in I \quad i \neq j \quad \forall k \in K,$$

where δ_k is the decay rate of asset type k (h^{-1}); t_{ijk} is the total transportation time of asset type k (h); $e^{-\delta_k t_{ijk}}$ is the quality of asset type k after the time t_{ijk} of transportation.

2.7 Refill Ports and Bunker Cost

Based on the results in existing studies [11, 13], we assume that the bunker cost $f(z_i)$ is a piecewise linear function with three linear segments about the bunker purchased

z_i. Hence, the single bunker cost at port i could be expressed as [17]:

$$f(z_i) = \begin{cases} C_i^{f0} z_i & 0 \le z_i \le W_i^1 \\ C_i^{f0}\left[W_i^1 + \lambda_i^1 \left(z_i - W_i^1\right)\right] & W_i^1 < z_i \le W_i^2 \\ C_i^{f0}\left[W_i^1 + \lambda_i^1 W_i^2 + \lambda_i^2 \left(z_i - W_i^2\right)\right] & W_i^2 < z_i \le W \end{cases}$$

where C_i^{f0} is the regular fuel price at port i, λ_i^1, λ_i^2 are the bunker fuel price discount factor for different refueling amounts at each port, $0 < \lambda_i^2 \le \lambda_i^1 \le 1$. W_i^1 and W_i^2 are the breakpoints of refueling amounts where liner shipping company can enjoy price discount, $W_i^1 \le W_i^2$. Hence, the total bunker cost of all ports is:

$$FC = \sum_{i=1}^{n} [F + f(z_i)] b_i.$$

where F is the fixed cost of vessel refueling, b_i is a binary variable used to decide whether to refuel at port i.

3 Model Formulation

This section presents a mixed integer non-liner mathematical model for the problem about vessel scheduling and refueling strategy of perishable assets with a collaborative agreement (VSRSPACA) in liner shipping.

3.1 Nomenclature

1. Sets

$I = \{1, 2, \ldots, n\}$	set of ports in the liner shipping route;
$K = \{1, \ldots, r\}$	set of perishable asset types;
$T_i = \{1, 2, \ldots, m_i\}, i \in I$	set of arrival TWs available at port i;
$S_{it} = \{1, 2, \ldots, g_{it}\}, i \in I, t \in T_i$	set of start times for TW t at port i;
$E_{it} = \{1, 2, \ldots, o_{it}\}, i \in I, t \in T_i$	set of end times for TW t at port i;
$H_{it} = \{1, 2, \ldots, w_{it}\}, i \in I, t \in T_i$	set of handling rates available at port i during TW t;
$P_{ijk} = \left\{1, \ldots, n_{ijk}\right\}, i, j \in I, i \neq j, k \in K$	set of all ports to which vessel is attached in the process of transporting asset type k with the port of origin i and the port of destination j.

2. *Parameters*

C^{oc} weekly vessel operational cost (USD/week);

$C_i^{lc}, i \in I$ unit late vessel arrival penalty at port *i* (USD/h);

C^{ic} unit container inventory cost (USD/TEU/h);

$C_i^{f0}, i \in I$ unit price of bunker fuel at port *i* (USD/ton);

$C_k^d, k \in K$ decay cost of asset type *k* (USD/TEU/%);

$C_{ith}^{hc}, i \in I, t \in T_i, h \in H_{it}$ handling cost at port *i* under handling rate *h* during TW *t* (USD/TEU);

$W_i^1, i \in I$ total bunker purchased at the first discount at the port *i* (tons);

$W_i^2, i \in I$ total bunker purchased at the second discount at the port *i* (tons);

$\lambda_i^1, i \in I$ bunker fuel price discount factor at port when the total bunker purchased in the range of $\left(W_p^1, W_p^2\right]$ (%);

$\lambda_i^2, i \in I$ bunker fuel price discount factor at port when the total bunker purchased exceeds W_i^2 (%);

$L_i, i \in I$ length of voyage leg *i* (between consecutive ports *i* and $i+1$, nmi), where, L_n is the length between consecutive ports *n* and 1 (nmi);

$Q_i, i \in I$ total amount of containers transported at voyage leg *i* (TEUs);

$D_i, i \in I$ total amount of containers handled at port (TEUs);

$R_{ijk}, i, j \in I, k \in K$ amount of asset type transported from the origin ports *i* to the destination ports *j* (TEUs);

$\delta_k, k \in K$ decay rate of asset type *k* (%/h);

$V^{\min}$ minimum vessel sailing speed (knots);

$V^{\max}$ maximum vessel sailing speed (knots);

V_D design vessel sailing speed (knots);

F_D bunker consumption of vessels at Design Speed (ton/day);

W vessel bunker tanker maximal capacity (tons);

F fixed cost of vessel refueling (USD);

I_0 minimum fuel inventory at the beginning (tons);

$HP_{ith}, i \in I, t \in T_i, h \in H_{it}$ handling productivity for handling rate *h* at port *i* during TW *t* (TEUs /h);

$TW_{its}^S, i \in I, t \in T_i, s \in S_{it}$ value of start time *s* for TW *t* at port *i* (h);

$TW_{ite}^E, i \in I, t \in T_i, e \in E_{it}$ value of end time *e* for TW *t* at port *i* (h).

3. *Decision Variables*

$v_i, i \in I$ vessel sailing speed at voyage leg *i* (knots);

$b_i, i \in I$ $= 1$, if the vessel is refueled while arriving at port *i* ($= 0$ otherwise);

$x_{it}^{TW}, i \in I, t \in T_i$ = 1 if TW t is selected at port i (= 0 otherwise);

$x_{its}^{S}, i \in I, t \in T_i, s \in S_{it}$ = 1, if start time s is selected for TW t at port i (= 0 otherwise);

$x_{ite}^{E}, i \in I, t \in T_i, e \in E_{it}$ = 1 if end time e is selected for TW t at port i (= 0 otherwise);

$y_{ith}, i \in I, t \in T_i, h \in H_{it}$ = 1 if handling rate h is selected at port i during TW t (= 0 otherwise).

4. *Auxiliary Variables*

m number of vessels to be deployed for service of the liner shipping route (vessels);

$t_i^a, i \in I$ arrival time at port i (h);

$t_i^b, i \in I$ vessel handling time at port i (h);

$t_i^d, i \in I$ vessel departure time from port i (h);

$t_i^w, i \in I$ waiting time of a vessel at port i (h);

$t_i^l, i \in I$ vessel late arrival at port i (h);

$t_i, i \in I$ sailing time of a vessel at port i (h);

$z_i^1, i \in I$ amount of bunker in the bunker tank when the vessel arrives at port i before taking bunker (tons);

$z_i^2, i \in I$ amount of bunker in the bunker tank when the vessel arrives at port i after taking bunker (tons);

$z_i, i \in I$ total bunker purchased at a port i (tons);

$f(z_i), i \in I$ bunker cost when total bunker purchased is z_i at port i;

$g(v_i), i \in I$ bunker consumption at voyage leg i when sailing at speed v_i;

$t_{ijk}, i, j \in I, k \in K$ total transportation time of asset type k from the origin port i to the destination port j (h);

$h(t_{ijk}), i, j \in Ii \neq j$ total change in quality of asset type k from the origin port i to the destination port j.

3.2 Model of VSRSPACA

A mixed integer nonlinear programming model (that will be further referred to as VSRSPACA) can be formulated as follows.

$$\min\{C^{oc}m + \sum_{i\in I}[F + f(z_i)]b_i + \sum_{i\in I}\sum_{t\in T_i}\sum_{h\in H_{it}} C_{ith}^{hc} D_i y_{ith} + C^{ic}\sum_{i\in I} t_i Q_i + \sum_{i\in I} C_i^{lc} t_i^l + \sum_{i\in I}\sum_{j\in I:j\neq i}\sum_{k\in K} C_k^d R_{ijk} h(t_{ijk})\} \tag{1}$$

$$0.1 b_i W \le z_i \le b_i W \quad \forall i \in I \tag{2}$$

$$z_i = z_i^2 - z_i^1 \quad \forall i \in I \tag{3}$$

$$z_1^1 = I_0 \tag{4}$$

$$z_i^1 \geq 0.1W \quad \forall i \in I \tag{5}$$

$$z_i^2 \leq W \quad \forall i \in I \tag{6}$$

$$z_{i+1}^1 = z_i^2 - g(v_i) \quad \forall i \in I, i < n \tag{7}$$

$$z_1^1 = z_n^2 - g(v_n) \tag{8}$$

$$\sum_{t \in T_i} x_{it}^{TW} = 1 \quad \forall i \in I \tag{9}$$

$$\sum_{t \in T_i} \sum_{s \in S_{it}} x_{its}^S = 1 \quad \forall i \in I \tag{10}$$

$$x_{its}^S \leq x_{it}^{TW} \quad \forall i \in I, \quad \forall t \in T_i, \quad \forall s \in S_{it} \tag{11}$$

$$\sum_{t \in T_i} \sum_{e \in E_{it}} x_{ite}^E = 1 \quad \forall i \in I \tag{12}$$

$$x_{ite}^E \leq x_{it}^{TW} \quad \forall i \in I, \quad \forall t \in T_i, \quad \forall e \in E_{it} \tag{13}$$

$$\sum_{t \in T_i} \sum_{h \in H_{it}} y_{ith} = 1 \quad \forall i \in I \tag{14}$$

$$y_{ith} \leq x_{it}^{TW} \quad \forall i \in I, \quad \forall t \in T_i, \quad \forall h \in H_{it} \tag{15}$$

$$t_i = \frac{L_i}{v_i} \quad \forall i \in I \tag{16}$$

$$t_i^b = \sum_{t \in T_i} \sum_{h \in H_{it}} y_{ith} \frac{D_i}{HP_{ith}} \quad \forall i \in I, \quad \forall t \in T_i, \quad \forall h \in H_{it} \tag{17}$$

$$t_{ijk} = \sum_{s \in P_{ijk}} t_s^b + \sum_{s \in P_{ijk}: s \neq n_{ijk}} t_s \quad \forall i, j \in I, \quad i \neq j, \quad \forall k \in K \tag{18}$$

$$t_i^d = t_i^a + t_i^b + t_i^w \quad \forall i \in I \tag{19}$$

$$t_i^l \geq t_i^a - \sum_{t \in T_i} \sum_{e \in E_{it}} TW_{ite}^E x_{ite}^E \quad \forall i \in I \tag{20}$$

$$t_i^w \geq \sum_{t \in T_i} \sum_{s \in S_{it}} TW_{(i+1)ts}^S x_{(i+1)ts}^S - t_i^a - t_i^b - t_i \quad \forall i \in I, \quad i < n \tag{21}$$

$$t_n^w \geq \sum_{t \in T_1} \sum_{s \in S_{1t}} TW_{1ts}^S x_{1ts}^S - t_n^a - t_n^b - t_n + 168m \tag{22}$$

$$t_{i+1}^a = t_i^d + t_i \quad \forall i \in I, \quad i < N \tag{23}$$

$$t_1^a = t_n^d + t_n - 168 \text{ m} \tag{24}$$

$$\sum_{i \in I} \left(t_i + t_i^b + t_i^w \right) = 168 \text{ m} \tag{25}$$

$$V^{\min} \leq v_i \leq V^{\max} \quad \forall i \in I \tag{26}$$

$$x_{it}^{TW}, x_{its}^S, x_{ite}^E, b_i, y_{ith} \in \{0, 1\} \tag{27}$$

The objective function (1) of the VSRSPACA mathematical model is to minimize the total liner shipping route service cost incurred by one ship per cycle, where the first term is the vessel cost, which is composed of the following components: Ship capital cost, crew cost and ship auxiliary oil consumption cost, the last term is total asset decay cost due to the change in quality of asset type after transportation from the origin port to the destination port. Note that other costs (including total fuel consumption cost, total port handling cost, total container inventory cost and total late arrival penalty in turn) are also taken into account. Equations (2) and (3) calculate the total amount of bunker purchased and the range at a port respectively. Equation (4) represent the minimal fuel inventory at the beginning. Equation (5) ensures that bunker fuel inventory before every refueling has a certain minimum level. Equation (6) ensures that bunker fuel inventory after every refueling does not exceed the bunker fuel capacity of a vessel. Equations (7) and (8) define the relation between bunker consumed and bunker fuel inventory before and after every refueling. Equation (9) ensures that only one service TW should be selected from the available vessel arrival TWs at each port of call. Equations (10) and (11) ensure that only one start time must be requested by the liner shipping company for the selected TW at each port of call. Equations (12) and (13) ensure that only one end time must be requested by the liner shipping company for the selected TW at each port of call. Equations (14) and (15) ensure that the vessel will be served under the selected handling rate for the selected TW at each port of call. Equation (16) computes the sailing time of vessels at each voyage leg. Equation (17) computes the handling of vessels time under the selected handling rate at each port of call. Equation (18) computes the total transportation time (including the total handling time at port and the total sailing time) of perishable asset type *k* from the origin port to the destination port. Equation (19) computes the departure time of vessels at each port. Equation (20) estimates the late arrival time of vessels at each port. Equations (21) and (22) estimate the waiting time of vessels at each port before the service begin. Equations (23) and (24) compute the arrival time of vessels at each port. Equation (25) ensures that the weekly service at each port should be provided. Equation (26) indicates that certain range of the vessel sailing speed at each voyage leg. Equation (27) is a binary constraint.

4 Algorithm

According to existing research literature [12, 18, 19] Using the "big M" piecewise linear secant approximations method ("M" is a big number), the mathematical model of VSRSPACA can be transformed into the mathematical model of VSRSPACAL, which can be solved efficiently by using off-the-shelf MILP solvers (such as CPLEX, etc.).

5 Numerical Experiments

This section will select a real liner shipping route to present some numerical experiments, in order to reveal some management insights.

5.1 Input Data Description

This study selected the Asia–North Europe AEU6 liner shipping route, which is served by the CHINA COSCO SHIPPING GROUP liner shipping company and is presented in Fig. 2.

The data from the available liner shipping literature [19–22] was used to generate the parameter values necessary for computational experiments.

5.2 Managerial Insights

Based on the above data, this paper first generated 1000 scenarios according to the above rules, and then used ILOG CPLEX 12.6 software to carry out numerical analysis on a computer with Pentium (R) i5 3.10 GHz memory of 4 GB. The results were obtained including: (1) Arrival time—***AT***; (2) Departure time—***DT***; (3) vessel sailing speed—***VS***; (4) Fuel consumption—***FC***; (5) Fuel inventory while arriving

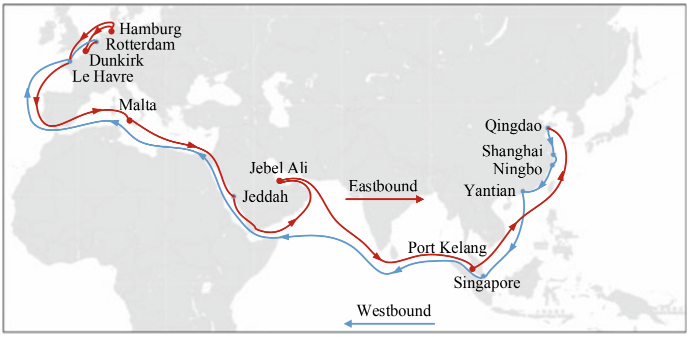


Fig. 2 AEU6 route

at the port—***FIA***; (5) Bunkered amounts—***BA***. The ship arrival/departure time and refueling strategy for a round-trip voyage are shown in Table 1.

It can be seen from Table 1 that the total time of a round-trip voyage is 1344 h. Therefore, the ship schedule of the liner route can be designed as 1344 h (8 weeks), and the number of ships allocated is 8. The vessel made three refueling decision at different ports, including Shanghai port, Singapore port and Malta port. Because bunkering policies are developed based on fuel prices at different ports and inventory levels in the tank, most bunkering takes place at ports where fuel prices are relatively low, and bunkering amount at these bunkering ports vary. This numerical experiment helps to determine the impact of bunkering policies, taking into account

Table 1 Results pertaining to the ship route AEU6

Port	*AT* (h)	*DT* (h)	*VS* (knots)	*FC* (tons)	*FIA* (tons)	*BA* (tons)
Qingdao	0	30	24.58	131	1000	–
Ningbo	47	74	24.66	42	869	–
Shanghai	80	106	24.59	253	827	538
Yantian	140	160	24.47	433	1112	–
Singapore	220	248	24.08	2386	679	3287
Le Havre	582	606	24.55	75	1580	–
Rotterdam	617	650	24.62	43	1505	–
Dunkirk	656	670	24.80	122	1462	–
Hamburg	685	700	24.71	154	1340	–
Le Havre	720	735	24.84	665	1186	–
Malta	822	841	24.72	510	521	3565
Jeddah	908	933	24.78	730	3576	–
Jebel Ali	1028	1063	24.89	1055	2846	–
Port Kelang	1199	1233	24.02	791	1791	–
Qingdao	1344	1374	–	–	–	–

the container handling rates efficiency and the choice of the vessel arrival window while determining bunkering ports and bunkering amount.

6 Conclusions and Future Research Extensions

Based on the above analysis, the following research conclusions and management implications can be obtained:

1. In the shipping service routes of ocean container liner with large reefers load, the liner shipping company should not only pay attention to the changes in transportation demand and bunker fuel price, but also further consider the structural characteristics of transportation demand (especially the perishable varieties and their decay rate characteristics) when making shipping scheduling and refueling strategy decisions. The liner shipping company shall take into account the differences in fuel prices (including discounts) between the ports of call in order to develop the optimal schedule design plan to minimize the total route service cost and ensure fuel supply.
2. The total asset decay cost and the total route service cost are positively correlated with bunker fuel prices. Specifically, as bunker fuel prices rise, the vessel sailing speed will decrease and cargo sailing times will be prolonged, leading to increase total port handling cost, container inventory cost and total late arrival penalty. The liner shipping company adopts the combination of schedule design and refueling strategy to effectively slow down the rising trend of the total route service cost under high bunker fuel prices.
3. The duration of the TWs has an important influence on the port handling rates selection and the time of vessel at port. The duration of the TWs is closely related to the total asset decay cost and even the total route service cost. The larger the duration of the TWs, the more flexibility the shipping company has to choose the appropriate port handling rates and ultimately reduce the total route service cost. Therefore, the liner shipping company should strengthen the close cooperation with port terminal operators on the route, sign the mutually beneficial and win-win agreement, in order to reduce the total route service cost and improve the revenue of the liner shipping company.

The analysis results of the numerical experiments verified the effectiveness and applicability of the model and solution proposed in this paper. It should be pointed out that this paper mainly studies the design of vessel schedule and refueling strategy of perishable assets when the liner shipping company is equipped with the same type of ships on a single route. The next research can consider the design of vessel schedule and refueling strategy of perishable goods under the condition of multi-route and multi-ship type.

Acknowledgements This study is supported by the National Natural Science Foundation of China (No. 71372088, 71871036), The Fundamental Research Funds for the Central Universities of China (No. 3132019023).

References

1. A. Rong, R. Akkerman, M. Grunow, An optimization approach for managing fresh food quality throughout the supply chain. Int. J. Prod. Econ. **131**(1), 421–429 (2011)
2. Z. Liu, Y. Yan, X. Qu, Y. Zhang, Bus stop-skipping scheme with random travel time. Transp. Res. C Emerg. Technol. **35**, 46–56 (2013)
3. M.G. Karlaftis, K. Kepaptsoglou, E. Sambracos, Containership routing with time deadlines and simultaneous deliveries and pick-ups. Transp. Res. E Logistics Transp. Rev. **45**(1), 210–221 (2009)
4. I. Norstad, K. Fagerholt, G. Laporte, Tramp ship routing and scheduling with speed optimization. Transp. Res. C Emerg. Technol. **19**(5), 853–865 (2011)
5. J.G. Rakke, M. Stålhane, C.R. Moe, M. Christiansen, H. Andersson, K. Fagerholt, I. Norstad, A rolling horizon heuristic for creating a liquefied natural gas annual delivery program. Transp. Res. C Emerg. Technol. **19**(5), 896–911 (2011)
6. M.A. Dulebenets, Minimizing the total liner shipping route service costs via application of an efficient collaborative agreement. IEEE Trans. Intell. Transp. Syst. **20**(1), 123–136 (2019)
7. P. Amorim, M.A.F. Belo-Filho, F.M.B. Toledo, C. Almeder, B. Almada-Lobo, Lot sizing versus batching in the production and distribution planning of perishable goods. Int. J. Prod. Econ. **146**(1), 208–218 (2013)
8. T.E. Notteboom, The time factor in liner shipping services. Marit. Econ. Logistics **8**(1), 19–39 (2006)
9. H.B. Bendall, A.F. Stent, A scheduling model for a high speed containership service: a hub and spoke short-sea application. Int. J. Prod. Econ. **3**(3), 262–277 (2001)
10. M. Qiang, W. Shuaian, H. Andersson, K. Thun, Containership routing and scheduling in liner shipping: overview and future research directions. Transp. Sci. **48**(2), 265–280 (2014)
11. D. Ronen, The effect of oil price on containership speed and fleet size. J. Oper. Res. Soc. **62**(1), 211–216 (2011)
12. M.A. Dulebenets, E.E. Ozguven, Vessel scheduling in liner shipping: modeling transport of perishable assets. Int. J. Prod. Econ. **184**, 141–156 (2017)
13. S. Wang, A. Alharbi, P. Davy, Liner ship route schedule design with port time windows. Transp. Res. C Emerg. Technol. 41, 1–17 (2014)
14. A. Alharbi, S. Wang, P. Davy, Schedule design for sustainable container supply chain networks with port time windows. Adv. Eng. Inform. **29**(3), 322–331 (2015)
15. S. Wang, Q. Meng, Robust schedule design for liner shipping services. Transp. Res. E Logistics Transp. Rev. **48**(6), 1093–1106 (2012)
16. J. Blackburn, G. Scudder, Supply chain strategies for perishable products: the case of fresh produce. Prod. Oper. Manage. **18**(2), 129–137 (2009)
17. S. Wang, Q. Meng, Liner ship route schedule design with sea contingency time and port time uncertainty. Transp. Res. Part B-Methodol. **46**(5), 615–633 (2012)
18. J.P. Vielma, S. Ahmed, G.J.O.R. Nemhauser, Mixed-integer models for nonseparable piecewise-linear optimization: unifying framework and extensions. Oper. Res. **58**(2), 303–315 (2010)
19. M.A. Dulebenets, A comprehensive multi-objective optimization model for the vessel scheduling problem in liner shipping. Int. J. Prod. Econ. **196**, 293–318 (2018)
20. Z. Yao, S.H. Ng, L.H. Lee, A study on bunker fuel management for the shipping liner services. Comput. Ind. Eng. **39**(5), 1160–1172 (2012)

21. C. Wang, J. Chen, Strategies of refueling, sailing speed and ship deployment of containerships in the low-carbon background. Comput. Ind. Eng. **114**, 142–150 (2017)
22. S. Wang, S. Gao, T. Tan, W. Yang, Bunker fuel cost and freight revenue optimization for a single liner shipping service. Comput. Ind. Eng. **111**, 67–83 (2019)

Research on the Common Delivery Model of Express Logistics in Urban and Rural Areas—A Case Study on Lujiang County, Hefei City

Zhou Yao, Shuihai Dou, Guanyi Liu, and Yanping Du

Abstract In the distribution environment at the end of the urban and rural areas, due to the scattered demand for urban and rural express delivery, there are problems such as long vehicle travel paths, high empty load rates, and long delivery times. In order to solve the problem that the current urban-rural regional end distribution is more difficult, a urban-rural regional distribution model based on the common distribution model is proposed. By integrating some urban and rural areas, they are regarded as a whole, and a common distribution center is established to replace the traditional county-level distribution center. Finally, using Lujiang County of Hefei as a case, a case simulation was conducted through MATLAB, which proved the effectiveness and feasibility of the common distribution model in urban and rural areas.

Keywords Common delivery · Urban and rural logistics · Courier end delivery

1 Introduction

As China's economic development enters a new era, the vigorous development of urban and rural logistics is of great significance for achieving urban-rural integration, promoting supply-side structural reform and optimizing industrial development. With the in-depth advancement of the construction of beautiful villages and the entry of e-commerce into villages, poverty alleviation work continues to rise, the growth rate of rural online retail sales continues to accelerate, and the results of e-commerce poverty

Z. Yao
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 13011151625@163.com

S. Dou (✉) · G. Liu · Y. Du
School of Mechanical and Electrical, Beijing Institute of Graphic Communication, Beijing, China
e-mail: doushuihai@126.com

G. Liu
e-mail: 15501092112@163.com

Y. Du
e-mail: duyanping@bigc.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_41

alleviation have been shown. According to data tracking, in the first half of 2019, rural online retail sales reached 777.13 billion yuan, a year-on-year increase of 21.0%, and the growth rate was 3.2% points higher than that of the country [1]. As we all know, with the rise of online shopping, e-commerce companies have gradually adopted the rural blue ocean market as an important strategic component for the next step. Rural e-commerce has become a new commercial market. However, in the construction of the rural logistics system, geographic environment, economic development factors, population density and other reasons have caused many problems to be solved at the same time, such as the distribution center is distributed, and the delivery time cannot be guaranteed. The "last mile" problem is particularly prominent. The improvement of rural logistics network nodes makes the orderly collection and distribution of "last mile" possible. At the same time, it can also effectively reduce urban and rural transportation costs, and make local people feel the happiness brought about by the rapid development of express logistics. Under this condition, urban–rural logistics under the common distribution model is a good way to solve the "last mile" of rural logistics.

Common Delivery is a kind of distribution task including integration of multiple manufacturers, operators or users, and intensive joint distribution through sharing or other methods to achieve the purpose of reducing costs and improving the efficiency of terminal logistics [2]. In short, it is a distribution activity carried out by some companies. Japan is the earliest place of common distribution. Japan has long learned that common distribution is an important way to solve the contradictory relationship between logistics service level and logistics cost. In the eyes of Yuasa and Yufu, the common distribution concept is: in order to reduce the empty load rate of the distribution vehicle, the distribution of a smaller distribution volume of a single company will be distributed to multiple companies with a smaller distribution volume. Each company reached a more reasonable and unified distribution method for the overall distribution of the alliance [3]. Schone's introduction and summary of the reasons for the common distribution are mainly due to the following three reasons: frequent shipments have caused great pressure on the logistics level, and the disgusting competition between different logistics companies has led to the poor allocation of logistics resources and The expansion of logistics demand caused by bad use and the unreasonable expansion of production demand have created conditions for the birth of joint distribution [4]. Yamada, Taniguchi, etc. studied the transportation path model in logistics distribution, combined co-distribution with the location selection and vehicle route planning model of the distribution center in traditional distribution, and finally proved the feasibility and effectiveness of co-distribution [5]. Yan [6] of Jiangsu University takes small and medium-sized enterprises as the research object, analyzes the related problems of profit distribution among small and medium-sized enterprises under the common distribution model, and constructs a common distribution benefit distribution model for small and medium-sized manufacturing enterprises. Zheng constructed a common distribution cost allocation model, applied to express cities, and analyzed the related impact factors [7]. Li from the perspective of game theory innovative morality proposed a common distribution model established in small and medium-sized cities below the third line, and established its construction cost model and operating cost model [8]. Xiao took rural enterprises as the research object and established the common delivery area and benefit distribution model of rural express delivery enterprises under the common delivery model [9]. Bao, Xing Researched the port solid waste green logistics system based on AHP and evaluated it [10].

According to the analysis of domestic and foreign research, we have done some research and analysis on common delivery at home and abroad. Most of the research on common delivery is based on the delivery model and delivery process, as well as the benefit distribution model [11],12. There is relatively little research on terminal distribution points and joint distribution between urban and rural areas. The research object is mainly focused on such as urban co-distribution, inter-city co-distribution, and the overall research of the logistics industry, and the research on co-distribution under the e-commerce express logistics environment in urban and rural areas is relatively rare. Based on the current development status of e-commerce logistics rural terminal outlets, this paper proposes a new type of common delivery model for urban and rural express delivery, which provides certain theoretical support for the establishment of a more scientific and efficient express logistics bottom delivery network.

2 Delivery Mode Overview

2.1 Third-Party Delivery Model

The so-called third party is the third party between the logistics service consignor (first party) and the consignee (second party). The third party delivery simply means that the consignor will hand over part or all of the goods to the consignee. Enterprises other than people have full power to entrust delivery.

With the future trend thinking about the development of supply chain integration, more and more companies will focus on their main business, outsourcing logistics, warehousing and other non-main business to third-party companies. The advantage is that it can increase the resources of the liberated enterprises in the logistics department, and allow the enterprises to invest their resources in the main business, thereby improving the core competitiveness.

2.2 Self-operated Delivery Model

Contrary to the third-party distribution model, the self-operated distribution model is that the enterprise establishes its own distribution network and completes the logistics distribution tasks of various enterprises through its own organization. The advantage of this model is that it enables enterprises to control their own resources and achieve integrated management, which can ensure service quality and expand enterprise business. The disadvantage is that the investment is high, there are certain requirements for the scale of the enterprise, and it can have enough energy, resources and funds to establish a distribution center and warehouse. This method is currently used by companies such as JD.com.

2.3 Interoperable Delivery Model

The so-called interoperable distribution is simply a way for certain enterprises to borrow each other's logistics resources and logistics configurations and use each other's logistics systems through certain agreements. The advantage is that enterprises do not need to invest a lot of time, materials and manpower to build a logistics system, and can expand the scope of their own logistics services. However, Hu Yong's distribution also has certain requirements for enterprises. It requires a high level of management among various enterprises and a strong organization and coordination ability.

2.4 Common Delivery Model

The so-called joint distribution is simply a combination of enterprises, overall planning, integration of logistics resources and facilities, unified allocation of logistics needs, unified scheduling of vehicles, and unified planning of distribution business. There are two modes, one is that a logistics company integrates the needs of multiple logistics companies in a unified manner, and arranges time and vehicles as a whole. Second, the companies mix cargos with the same batch of vehicles in the distribution process in response to logistics needs.

3 Common Delivery Model in Urban and Rural Areas

3.1 Realization Path of Urban-Rural Common Delivery

1. Form a joint express delivery enterprise alliance. In order to complement resources between express companies, reduce the construction cost of logistics resources and facilities and equipment, and improve the level of logistics management. After the express enters the urban and rural areas, the urban and rural express co-distribution enterprise alliances formed by the express companies jointly deliver.

2. Build express sharing network information platform. The joint delivery of urban and rural express delivery is aimed at obtaining the large-scale benefits of express delivery, based on advanced technical means, and sharing the information of express delivery business as a link, so as to build a coordinated operation mode, and finally achieve information sharing and collaborative operation.
3. The use of co-express delivery forecasting technology. After the courier recipient places an order on the e-commerce website or enterprise, the courier is sent by the e-commerce website, and the logistics information of the courier will be transmitted to the urban and rural express shared network information platform. With the order information and logistics information, the urban and rural express shared network information platform can use this as a basis to predict which couriers will soon arrive at the urban and rural express co-distribution center, and effectively integrate the resources of the courier with the same recipient who will arrive, thus Reaching the goal of express recipients receiving couriers from different courier companies at the same time.

3.2 Working Mode of Regional Common Distribution Center

Upstream and downstream are divided into four levels, from top to bottom for each express company, comprehensive sorting center, regional common distribution center and outlets. The express mail sent from a courier company is finally delivered to the target customer, forming a forward logistics chain. The operating mode is shown in Fig. 1.

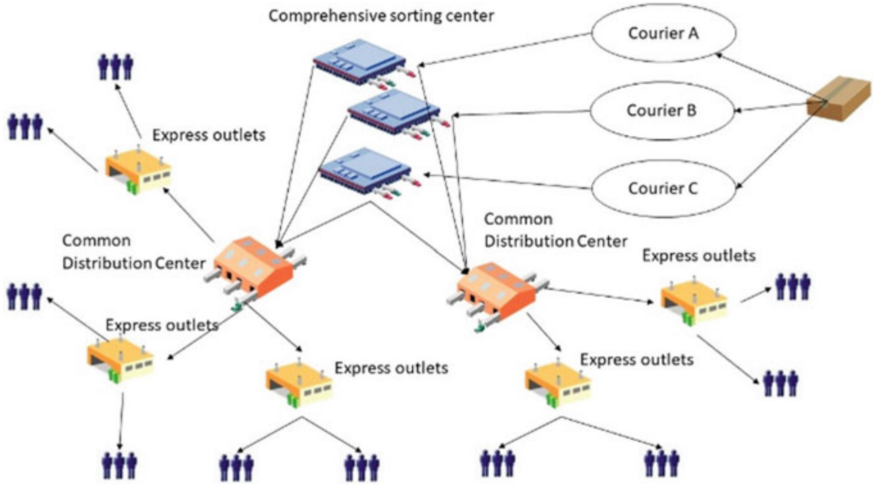


Fig. 1 Express co-distribution delivery mode

Comprehensive sorting centers are generally located in prefecture-level cities, usually as an important logistics node of a province. The common distribution center in the urban and rural areas is different from the general situation. It is not located in the county town of the prefecture-level city, but in some townships in the county town. In this model, a county usually has several distribution centers in urban and rural areas. They divide the towns of the county into several areas, and are responsible for the distribution tasks in their respective areas. In this model, the destination is that the express delivery of the township passes through the comprehensive sorting center. The destination is not the distribution center of the county, but the regional common distribution center corresponding to the township.

The couriers of different companies go through their respective comprehensive sorting centers and are sorted to a common distribution center in the urban and rural areas. The distribution center does not sort on the basis of the courier branch company, but sorts on the basis of the delivery location of the courier. Then it will be transported to the terminal distribution network according to the distribution address, and finally the network will deliver the shipment to the user.

The package delivery process of the common distribution model is shown in Fig. 2, and the process is similar to dispatching. Express delivery is collected through the outlets, and the outlets are transported to the urban and rural regional common distribution center, and then the regional common distribution center sorts according to different express companies, dispatches to the upper-level sorting center, and finally sorts and transports the sorting center. If it is the same city express, it will be sorted out separately in the regional distribution center and sent to the outlet or the distribution center in the same city.

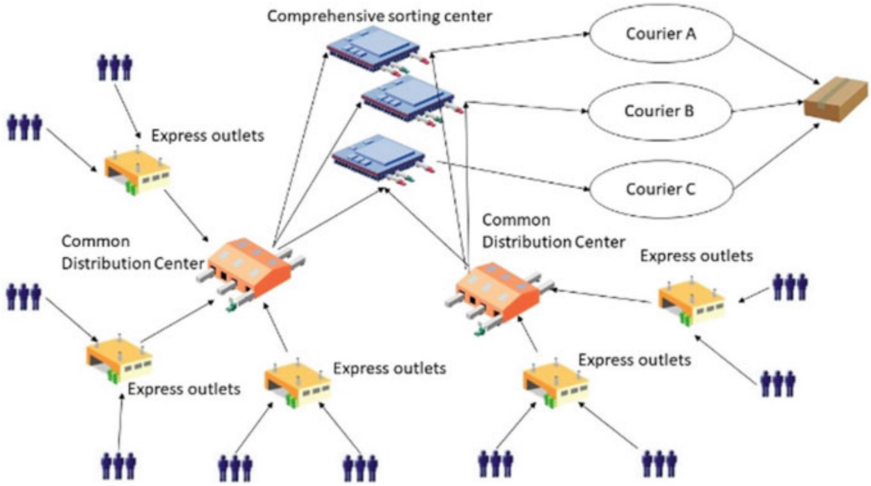


Fig. 2 Courier common delivery package mode

4 Case Simulation of Lujiang County, Hefei City

4.1 Case Assumption

In this case, it is assumed that the vehicle load is only affected by the volume, and the weight factor is ignored. Under the premise of co-distribution by various express delivery companies, considering that the population of the county town of Lujiang County accounts for one third of the total population of Lujiang County, the density of express delivery demand generated by the county town is significantly higher than that of other townships. Therefore, in the design of the common delivery plan for express delivery in Lujiang County, Lucheng Town (urban area) is divided from 16 other townships, and a comprehensive express service center is set up in 16 townships such as Nihe Town and Ketan Town, and express self-pickup services are configured point. The delivery services of the six express delivery companies of Yuantong (YT), Zhongtong (ZT), Yunda (YD), Shentong (ST), Best (BS) and SF Express have been integrated. Lujiang County's distribution network is tentatively unchanged and the existing express network is still used as an example. Since Lujiang County has three main roads, G330, G346, and S319, passing through Lujiang County Town, the remaining townships in Lujiang County are divided into three parts. Therefore, in this case, it is assumed that 16 townships are divided into three categories, and the Lujiang County City Express Network is established separately A distribution center.

In order to verify the correctness and feasibility of the fuzzy improved artificial potential field method proposed in this paper, the static obstacle environment and dynamic obstacle environment are constructed by using MATLAB software, and the traditional artificial potential field method and the fuzzy improved artificial potential field method are respectively used for road strength planning simulation experiment.

4.2 Data Research and Analysis

1. *Express location data.* Due to the different delivery ranges and inconsistent number of outlets between different express companies, at the same time, the number of express outlets in Lucheng Town accounts for one-third of the total number of express outlets in Lujiang County, and different outlets in the same township operate daily when they are operated separately. The amount of dispatched parts is low, so the data in this part is simplified. The data of Lucheng Town's outlets remain unchanged, and the integration of community express delivery outlets in each township is integrated, which means that each community will build a courier common integrated self-lifting center. The data of each township network in this plan is shown in Table 1.

   After visiting and investigating, we collected the address information of the distribution centers of YT, ZT, YD, ST, BS and SF in Lujiang County. The results are shown in Table 2.

Table 1 Longitude and latitude coordinates of township network

	Name	Longitude	Latitude
1	PG	117.409877	31.22822
2	BH	117.464826	31.21919
3	LQ	117.446519	31.15111
4	HT	117.50751	31.11317
5	FS	117.433179	31.10644
6	LD	117.456472	31.051116
7	DQ	117.360299	31.0134
8	LH	117.293897	31.01655
9	NH	117.327108	31.09041
10	LQ	117.190511	31.14543
11	GM	117.119743	31.21121
12	KT	117.148776	31.24257
13	CB	117.181546	31.22708
14	TC	117.100389	31.340118
15	WS	117.189923	31.33701
16	CG	117.217061	31.307898
17	SQ	117.50556	31.398833
18	BS	117.39011	31.471916
19	TD	117.271481	31.50513
20	ST	117.278407	31.427137
21	JN	117.221495	31.39952
22	GH	117.166348	31.43009

Table 2 Address coordinates of YT, ZT, YD, ST, BS and SF Distribution Center

	Name	Longitude	Latitude
1	YT	117.317669	31.247587
2	ZT	117.317669	31.247587
3	YD	117.317669	31.247587
4	ST	117.279406	31.237135
5	BS	117.317669	31.247587
6	SF	117.317669	31.247587

2. *Delivery volume and delivery size survey.* According to the interview and investigation, the average daily express number of Lujiang County township community (except urban area) is shown in Table 3.

Table 3 Average daily express delivery volume of various township communities in Lujiang County

	Name	Number		Name	Latitude
1	PG	1500	12	KT	640
2	BH	1430	13	CB	860
3	LQ	1230	14	TC	2400
4	HT	550	15	WS	330
5	FS	600	16	CG	850
6	LD	500	17	SQ	2650
7	DQ	420	18	BS	2100
8	LH	960	19	TD	1000
9	NH	2480	20	ST	1450
10	LQ	1850	21	JN	1590
11	GM	300	22	GH	1230

At the same time, we randomly sampled 110 pieces of couriers delivered daily by the express outlets of Lujiang County to calculate the average volume. The estimated express size is 4383.536cm^3.

3. *Distribution vehicle parameter survey and vehicle capacity estimation.* In each express delivery center, there are a number of delivery vehicles. This vehicle is the JAC Suzuka E series, with a load capacity of 1.8 T, a cargo volume of 13.6 m^3, a mailbox of 100 L, and a displacement of 2.7 L. Using No. 0 diesel, the diesel oil price is 5.13 yuan / L, one liter of diesel can be used for the vehicle to travel about 8.5 km, then the vehicle costs 0.6 yuan for 1 km. After a single refueling, the maximum driving distance is 850 km, and the cost of renting the vehicle for one day is about 150 yuan.

According to the vehicle capacity of 13.6m^3, assuming a full load factor of 95%, the loaded vehicle capacity is 2947.39, and the integer is approximately estimated to be 2947. That is to say, a JAC car Suzuka E series has a load capacity of 1.8 T and a cargo volume of 13.6 m^3. The number of couriers can be loaded at a time is 2937.

4.3 Result Analysis

1. *k-means clustering algorithm for distribution center location.* The *K*-means clustering algorithm is used to divide the township network points (divided into three categories) for operation, and the results are shown in Fig. 3.
   Circles indicate express outlets, asterisks indicate distribution centers, blue indicates first category, red indicates second category, and green indicates third category.
   The three common distribution centers are located at {117.3873, 31.4357}, {117.1634, 31.3330}, {117.3577, 31.1454}.

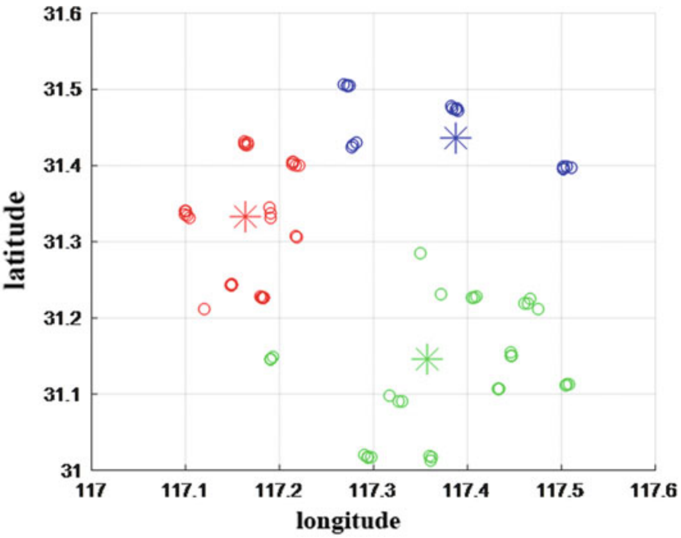


Fig. 3 Express vehicle models

The No.1 distribution center is located in BS, and undertakes the distribution tasks of BS, SQ, TD, and ST.

The No.2 distribution center is located in WS, and undertakes the distribution tasks of CG, KT, TC, WS, GH, JN, GM, CB.

The No. 3 distribution center is located in NH, and undertakes the distribution tasks of BH, LQ, FS, LH, NH, LQ, PG, HT, LD, DQ.

Since the distribution center No.3 is far from the community and the distance from Nihe Town is relatively short, and the transportation is convenient (the national highway G330 can be quickly reached), in actual situations, it may be considered to establish the distribution center No. 3 in NH, coordinates {117.3271, 31.0904}.

2. *Ant colony algorithm to solve VRP problem.* The results obtained by solving the distribution routes of the three distribution centers using ant colony algorithm are shown in the Figs. 4, 5 and 6.

The final delivery route results are shown in Table 4.

Under the improved common distribution model, 11 vehicles are needed, with a total mileage of 355.9 km, a total cost of 1706.34 yuan, and an average vehicle loading rate of 83.41%.

Using the ant colony algorithm to solve the ST, ZT, YT, SF, YD, BS distribution routes under the existing plan, the results are shown in the Figs. 7, 8, 9, 10, 11 and 12.

The delivery route results are shown in Table 5.

In summary, under the current circumstances, the total transportation distance of the six express delivery companies is 1630.4 km, requiring 11 vehicles, and the total cost is 2778.24 yuan. The average vehicle loading rate is 76.47%.

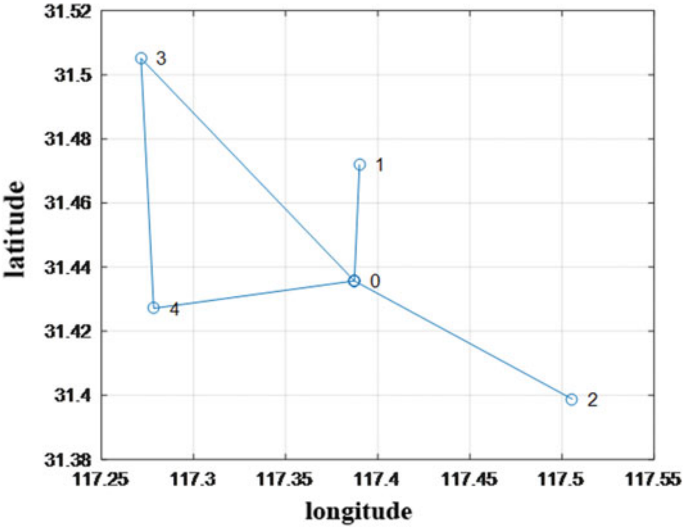


Fig. 4 Distribution route map of No. 1 distribution center

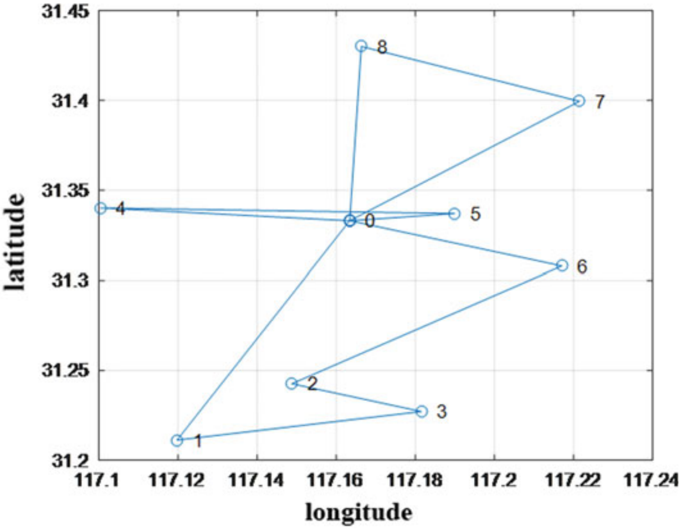


Fig. 5 Distribution route map of No. 2 distribution center

The difference between the common distribution model and the existing model is shown in Table 6.

It can be concluded that the common distribution model in urban and rural areas can greatly reduce the total distance traveled by vehicles, thereby reducing the cost

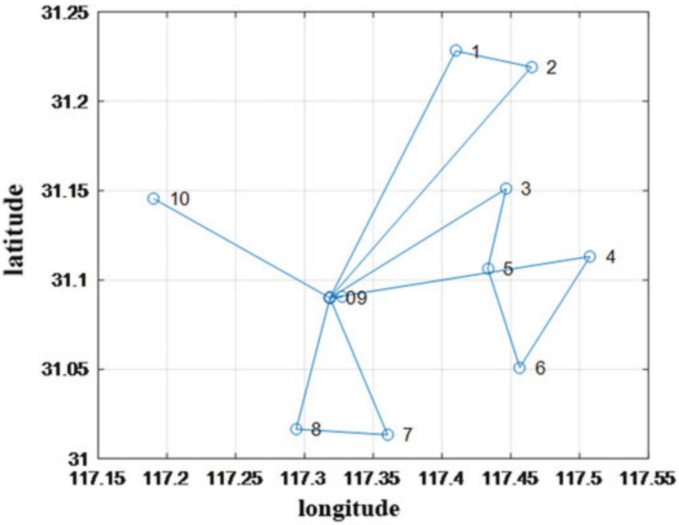


Fig. 6 Distribution route map of No. 3 distribution center

Table 4 Common delivery results

No. 1 distribution center	
Optimal distribution plan	0 > 1 > 0 > 2 > 3 > 4 > 0
Total cost	493.92 yuan
Delivery distance	73.2 km
Number of vehicles	3
No. 2 distribution center	
Optimal distribution plan	0 > 4 > 5 > 0 > 8 > 7 > 0 > 6 > 2 > 3 > 1 > 0
Total cost	501.9 yuan
Delivery distance	98.5 km
Number of vehicles	3
No. 3 distribution center	
Optimal distribution plan	0 > 10 > 0 > 8 > 7 > 0 > 1 > 2 > 0 > 3 > 5 > 6 > 4 > 0 > 9 > 0
Total cost	710.52 yuan
Delivery distance	184.2 km
Number of vehicles	5

of distribution. At the same time, it can also reduce the number of vehicles used and increase the vehicle loading rate and utilization rate.

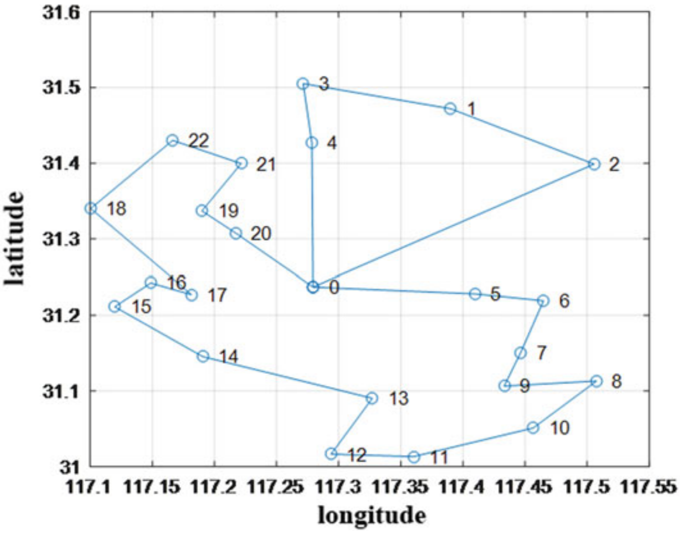


Fig. 7 ST delivery route

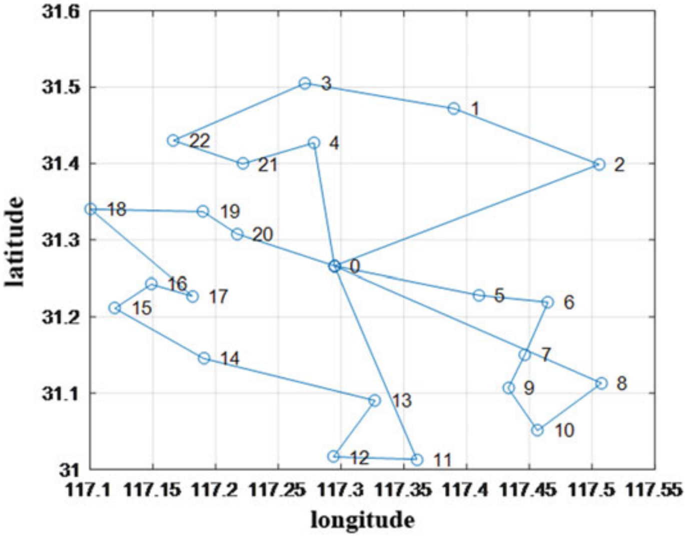


Fig. 8 ZT delivery route

5 Conclusions

In order to improve the delivery capacity and efficiency of the express delivery in urban and rural areas, a common distribution model in urban and rural areas is

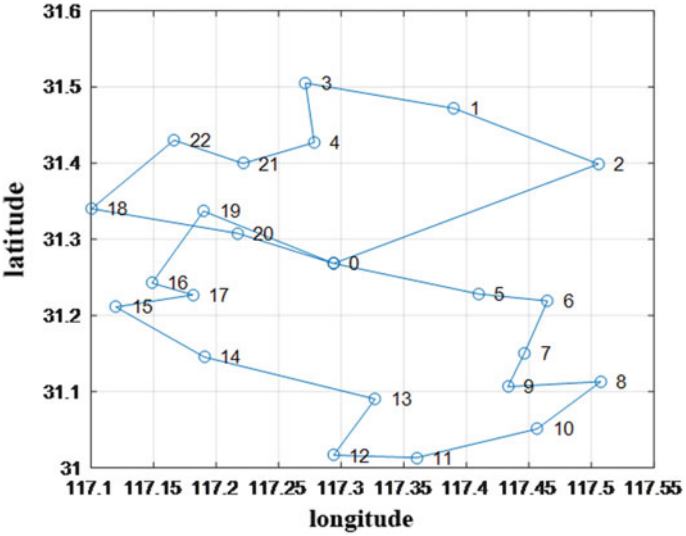


Fig. 9 YT delivery route

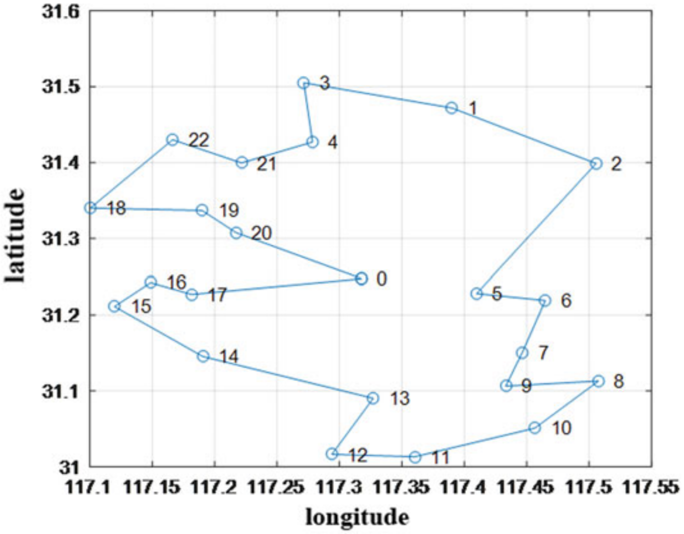


Fig. 10 SF delivery route

proposed, and its implementation path and its work mode of dispatching and package delivery are described. In order to solve the problem of time and cost related to the end of urban and rural distribution, by treating some towns as an urban and rural area, a common distribution center is set up in the urban and rural area to replace the traditional county-level distribution center. At the same time, using Lujiang County

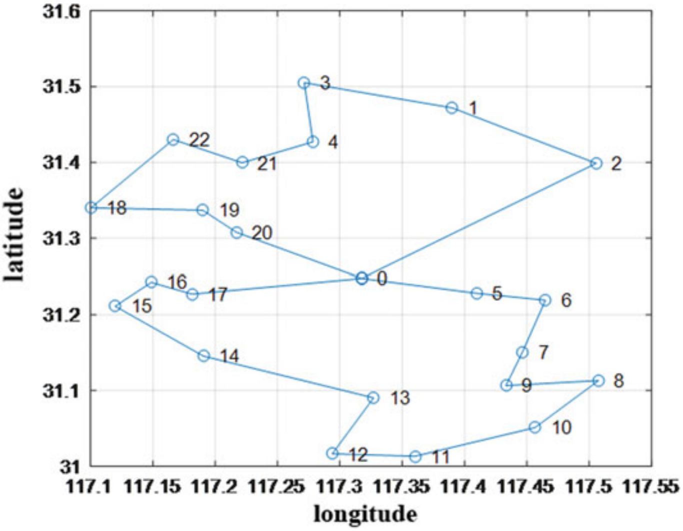


Fig. 11 YD delivery route

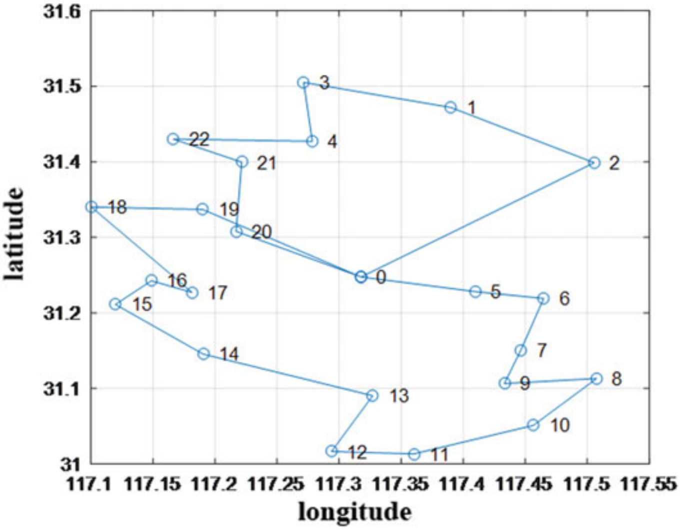


Fig. 12 BS delivery route

of Hefei as a case, taking 6 express companies of Shentong, Zhongtong, Yuantong, SF Express, Yunda and Bast as research objects, the distribution center location and vehicle were solved by using k-means clustering algorithm and ant colony algorithm respectively Delivery route. And compared with the existing distribution model to

Table 5 The existing plan delivery results

ST distribution center	
Total cost	463.2 yuan
Delivery distance	272 km
Number of vehicles	2
ZT distribution center	
Total cost	641.52 yuan
Delivery distance	319.2 km
Number of vehicles	3
YT distribution center	
Total cost	463.92 yuan
Delivery distance	273.2 km
Number of vehicles	2
SF distribution center	
Total cost	292.8 yuan
Delivery distance	238 km
Number of vehicles	1
YD distribution center	
Total cost	452.76 yuan
Delivery distance	254.6 km
Number of vehicles	3
No. 3 distribution center	
Total cost	464.04 yuan
Delivery distance	273.4 km
Number of vehicles	2

Table 6 The difference between the common distribution model and the existing model

	Common delivery	The existing model	Difference
Total cost	355.9 km	1630.4 km	78.17% reduction
Delivery distance	1706.34 yuan	2778.24 yuan	38.58% reduction
Number of vehicles	11	12	Reduce 1
average vehicle loading rate	83.41%	76.47%	6.94% increase

carry out example verification simulation, proved the effectiveness and feasibility of the common distribution model in urban and rural areas.

Acknowledgements This work was supported by the Scientific Research Program of Beijing Education Commission under Grant KM201910015004.

References

1. Total retail sales of rural consumer goods. Ministry of Commerce Data Center, 2019. [Online]. Available: https://data.mofcom.gov.cn/gnmy/societyCountry.shtml. Accessed: 01 July 2019
2. J.J. Zhang, T. Jiang, A review of research on common delivery of electronic commerce. China Market **19**(34), 87–90 (2013)
3. K. Yuasa, *Introduction to Logistics Management* (China Railway Press, Beijing, 1986).
4. A. Schone, W. Schimd, On the joint distribution of a quadratic and a linear form in normal variables. J. Multivar. Anal. **29**(02), 163–182 (2000)
5. T. Yamada, E.Taniguchi, Y. Itoh, Co-operative vehicle routing model with optional location of logistics terminals. City Logistics (02), 139–153, (2001)
6. L. Yan, Research on Common Distribution Mode of Small and Medium-Sized Manufacturing Enterprises. M.S. thesis, Jiangsu University (2010)
7. Z. Sheng, Research on Cost Allocation Model of Common Delivery Alliance in Express City. M.S. thesis, Chongqing University (2014)
8. T.T. Li, A Study on the Urban Common Distribution Model Dominated by Third-Tier and Below Urban Commercial Centers. M.S. thesis, Beijing Jiaotong University (2016)
9. R. Xiao, Research on Profit Distribution of Rural Express Enterprises Based on Common Delivery. M.S. thesis, Donghua University (2018)
10. X.X. Bao, X.C. Xing, Evaluation of green logistics system of solid waste at ports based on analytic hierarchy process. Environ. Eng. Manage. J. **18**(11), 2491–2499 (2019)
11. L.A. Ocampo, C.M. Himang, A. Kumar, M. Brezocnik, A novel multiple criteria decision-making approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy AHP for mapping collection and distribution centers in reverse logistics. Adv. Prod. Eng. Manage. **14**(3), 297–322 (2019)
12. G.I. Fragapane, C. Zhang, F. Sgarbossa, J.O. Strandhagen, An agent-based simulation approach to model hospital logistics. Int. J. Simul. Modell. **18**(4), 654–665 (2019)

Analysis of Lemon Company's Cross-Border E-Commerce Logistics Distribution Mode Selection

Li Qin Hu, Amit Yadav, Hong Liu, Sami Azam, Asif Karim, Bharanidharan Shanmugam, Abdul Hasib Siddique, and Mehedi Hasan

Abstract A detailed study of cross border e-commerce of lemon company has been done, it analyzes and summarizes its business products, main consumer objectives, existing logistical distribution model and combines the status of logistical operations of Lemon Company to analyze the logistical aspects of Lemon Company. Existing problems and factors that affect the choice of the logistical company's distribution model are analyzed in detail. An index system for the selection of cross-border e-commerce logistical distribution models has also been constructed. Existing logistical system has been comprehensively analyzed through expert scoring and other methods. Quantitative scoring of the index factors of the distribution model has been computed using the analytic hierarchy process and Matlab. Final comprehensive score of the three types of logistics distribution modes of the cross-border e-commerce company has been computed. A choice has been made for the logistical and distribution mode suitable for Lemon Company which can be used as overseas warehouse and distribution mode.

Keywords Cross-border e-commerce · Logistics distribution model · Cross-border logistics

L. Q. Hu
Department of Information Management, Chengdu Neusoft University, Chengdu, China

A. Yadav
Department of Information and Software Engineering, Chengdu Neusoft University, Chengdu, China

H. Liu
Department of Human Resources, Chengdu University of Technology, Chengdu, China

S. Azam · A. Karim (✉) · B. Shanmugam
College of Engineering, IT and Environment, Charles Darwin University, Darwin, NT, Australia
e-mail: asif.karim@cdu.edu.au

A. H. Siddique · M. Hasan
Department of Computer Science and Engineering, University of Science and Technology Chittagong, Chittagong, Bangladesh

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_42

1 Introduction

1.1 Lemon (L) Company Background

Lemon Cross-border e-commerce Company is an export cross-border B2C enterprise integrating Amazon trade export and logistics system services. It has its own warehouses in Shenzhen, Shanghai and other places, with overseas sales of about 16 million RMB/year. Company Lemon's export retail e-commerce platform has opened up various links such as logistics, payment, customs declaration, and has developed its own ERP system, the company is using franchise model to expand its business. At present, the number of franchise have reached 50 and all shipments are handled by the logistical department of Lemon Company [1].

Company's upstream supply chain includes well-known domestic manufacturers, and on the other hand downstream companies are mainly domestic or foreign logistical companies with cross-border logistical operations. At present, the company mainly serves C-end consumers mainly in USA, UK, and provides one-stop shopping services for foreign consumers. The flowchart of the companies (Fig. 1).

1. *Virtual delivery*:
   According to the characteristics of the goods and customer needs, selection of the appropriate logistics distribution mode is done.
2. *Prompt purchase*:
   The company finds the cargo information and notify the seller to send it quickly.
3. *Transit of goods*:
   After receiving the goods from the seller Lamon company does a QA survey of the product and sends it forward if it qualifies according to the standard in case if fails to comply to the standard then it notifies the seller to send a new product.
4. *Cargo tracking*:
   A real time tracking of the shipment is done through feedback method. Any abnormalities is also taken care off.

In a fiercely competitive market, as a small and medium-sized cross-border export e-commerce company must not only consider more supporting professional services and technological improvements in its development process, but also consider how to better serve the customers while making profit is optimal. Therefore, saving logistics costs and choosing the most reasonable logistics distribution model will become the profitable channel for Lemon Company. On the premise of ensuring the quality of logistics services, it is particularly important to choose a reasonable logistics distribution model.

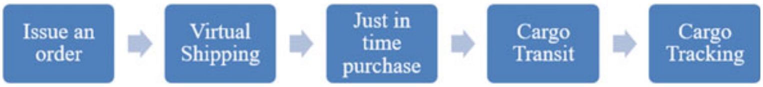


Fig. 1 Delivery process of Lemon Company. *Data Source* WWW.100EC.CN

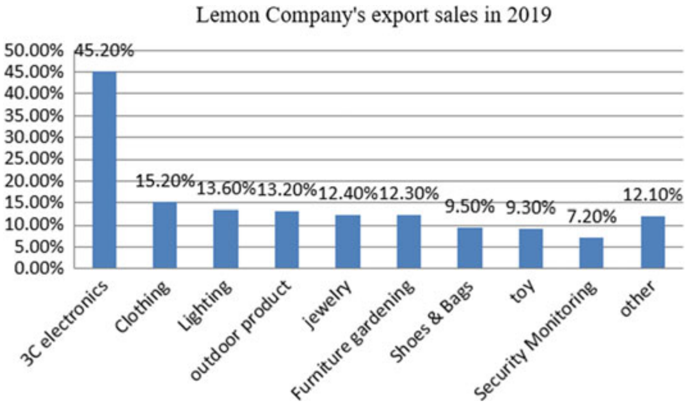


Fig. 2 Lemon Company's export sales in 2019. *Data Source* WWW.100EC.CN

1.2 Company L Export Product

The main products exported by Lemon Company are based on the daily need of a customer. For example, lighting, safety monitoring, clothing, These are divided into 3C electronic products and household products, which are all non-dumping products. 3C electronic products mainly include: mouse, keyboard, screen display, camera, game console, etc. Among these 3C products the sales of mouse and keyboard are relatively higher than the others. Clothes, jewelry, toys, clothing, hats, stationery are all products under the household category. The export sales of the Lemon Company in 2019 is shown in Fig. 2.

1.3 Overseas Operation

In order to avoid logistics congestion during the peak period, and to make sure that the goods can reach the consumers safely and quickly the company took some innovative measures.

At a fixed time every year, Lemon Company delivers the goods in advance to an overseas warehouse which is close to Nepal, thereby realizing the time value of storage [2]. Lemon Company then publishes its products on the cross-border e-commerce platform. Foreign buyers place orders on the that platform according to the demand. After receiving the order information, the sellers will share its shipping information and notify the overseas warehouse manager to send it in advance. Overseas warehouse staff will cooperate with overseas courier companies to securely deliver the products to overseas buyers. The advantage of this delivery model is that it saves time required for customs clearance and other processes. Hence one can responds to customer needs more quickly [1]. Although overseas warehousing and distribution modes are very advantageous in terms of logistical timeliness, but there is huge pressure from the same industry, other emerging logistics and distribution modes such as border warehouses.

For international express, Lemon Company uses the three international express delivery services namely TNT, UPS and DHL. Generally, the aging time from China to the United States is about 3–7 days. As international courier is fast, so Lemon Company usually chooses international courier to ship according to customers' requirements. In order to ensure customer satisfaction and on time delivery of the goods, Lemon Company chooses an international express logistics distribution model [1].

At present, postal parcel is the most common logistical distribution model of Lemon Company. The postal parcel delivery model is mainly used to deliver some home kitchen products. Although the delivery time is slower than other logistical delivery models, but there are no additional delivery fee for this mode to deliver in the remote areas.

2 Lemon Company's Cross-Border Logistics Distribution Model Selection

Even though the sales of the company is quite good but there are many areas where Lemon Company still needs to improve such as logistics cost issues, on time delivery of logistics and low number of returning customers. In order to solve these problems the following need to be established in the logistical distribution. Mode selection indicators for Lemon Company from both the strategic and tactical levels and comprehensively analyze the various indicators that affect the Lemon Company logistics distribution. In order to establish an analysis model Lemon Company's choice of logistics distribution mode needs to be more rational and quality of service needs to be high [1, 3].

According to the company's strategic and tactical index factors. Distribution can be divided into the different level of the company's logistical distribution mode. This is done according to the analytic hierarchy process. The second-level for strategic and tactical is the standard level. It includes export tax rate laws and regulations, consumer environment, logistics technology, product attributes, the price of logistics, the timeliness of logistics, the quality of logistics services, the strength of the company, and the needs of buyers are the sub-criteria layer, and the lowest layer is the program layer [2, 3] as shown in Fig. 3.

Company exports and the logistics distribution model it chooses to export. Of course, the products exported by Lemon Company also need to choose the appropriate logistics and distribution mode according to the attributes of the products and the consumer's requirements for timeliness. For the Company, the logistics price refers to the logistics costs required in the unit order, including the surcharges that may occur when selecting the international express logistics distribution mode, remote

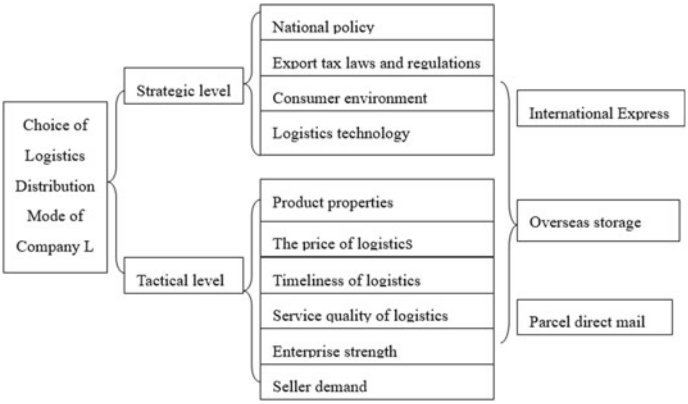


Fig. 3 Index system table. *Data Source* WWW.100EC.CN

area distribution fees, and transportation costs based on weight or volume, etc. In addition, the overseas storage costs, manual sorting, and packaging costs are collectively calculated as the company's logistics price. This is also the direct cost of Lemon Company's loss in a short period of time. The scale of the company will limit the Lemon Company's choice of logistics distribution mode, such as the idea of building overseas warehouses, can only rent overseas warehouses, relying on third parties [1, 3].

The index weights obtained by the grey correlation method combined with the analytic hierarchy process are used to calculate the comprehensive score. First, the calculated scoring level is established, and then analytic hierarchy method is similarly set as a nine-level scoring standard [2, 4, 5].

Secondly, for the above indicators, Lemon Company interviewed 15 experts in related fields to score and calculate their average scores, and constructed corresponding judgment matrices for the secondary and tertiary indicators. The related scoring tables can be summarized as shown in Table 1. According to expert evaluation the strategic level matrix and tactical level matrix had been created [2, 4, 5].

By using the data of the three tables, the weights of the secondary and tertiary indicators had been calculated using the Matlab matrix factory.

Next, from the 15 expert scoring tables, 5 copies were randomly selected using the gray correlation method to grade the index to construct an index sample matrix.

$$\mathrm{En}(\mathrm{n} = 1, 2, 3, 4, 5) : \text{Means an expert.}$$

Table 1 Expert evaluation table for secondary indicators

	Strategic level	Tactical level
Strategic level	1	1/3
Tactical level	3	1

Iij (i = 1,2; j = 1, 2, 3, 4, 5, 6): Represents the indicators that affect the selection of the logistics model of Lemon Company. The abbreviations of related indicators are I_{11} = National policy, I_{12} = Export tax laws and regulations, I_{13} = Consumer environment, I_{14} = Logistic technology, I_{21} = product properties, I_{22} = the price of logistics, I_{23} = Timeliness of logistics, I_{24} = Service quality of logistics, I_{25} = Enterprise strength, I_{26} = Seller demand. Experts' scores on international express delivery are shown in Table 2, scores of experts in direct mail mode is shown in Table 3 and scores of experts for overseas warehousing is shown in Table 4.

By using the above related score calculations have been done in the next section to obtain a comprehensive score of the three schemes, which provides a reference for the selection of the company's cross-border logistics distribution model.

Table 2 Expert scores for international express delivery models

I_1	E1	E2	E3	E4	E5	I_2	E1	E2	E3	E4	E5
I_{11}	6	7	6	8	7	I_{21}	5	4	6	6	4
I_{12}	8	7	7	8	7	I_{22}	6	6	7	6	7
I_{13}	6	5	7	6	6	I_{23}	4	5	5	4	6
I_{14}	7	7	9	7	8	I_{24}	3	5	4	3	4
						I_{25}	5	4	5	6	6
						I_{26}	5	7	5	6	5

Table 3 Scores of experts in direct mail mode

I_1	E1	E2	E3	E4	E5	I_2	E1	E2	E3	E4	E5
I_{11}	6	8	8	6	7	I_{21}	4	5	3	4	4
I_{12}	8	6	7	7	7	I_{22}	5	4	3	3	4
I_{13}	8	7	8	9	8	I_{23}	7	8	6	7	8
I_{14}	5	6	7	6	7	I_{24}	4	3	5	5	5
						I_{25}	4	3	3	5	3
						I_{26}	5	5	7	5	7

Table 4 Scores of experts for overseas warehousing

I_1	E1	E2	E3	E4	E5	I_2	E1	E2	E3	E4	E5
I_{11}	7	9	8	7	7	I_{21}	7	5	6	8	6
I_{12}	6	7	6	5	6	I_{22}	5	6	6	4	5
I_{13}	8	8	7	8	7	I_{23}	7	8	8	7	6
I_{14}	4	5	4	3	5	I_{24}	5	4	6	5	4
						I_{25}	4	4	6	5	3
						I_{26}	3	5	4	5	4

3 Construction of Evaluation Model and Analysis

The entire evaluation model calculation process is as follows:

3.1 Analytic Hierarchy Process to Calculate Index Weight

Through the analytic hierarchy model and the expert's scoring of indicators. Two-level indicator construction matrix, three-level indicator strategic-level indicator judgment matrix, and three-level tactical-level indicator judgment evidence matrix have been constructed [2, 4–6].

Secondary indicator judgment matrix

$$\begin{bmatrix} 1 & 1/3 \\ 3 & 1 \end{bmatrix}$$

Three-level indicator Strategic level indicator judgment matrix:

$$\begin{bmatrix} 1 & 4 & 3 & 3 \\ 1/4 & 1 & 1/5 & 1/3 \\ 1/3 & 5 & 1 & 3 \\ 1/3 & 3 & 1/3 & 1 \end{bmatrix}$$

Three-level indicator tactical level indicator judgment matrix:

$$\begin{bmatrix} 1 & 1 & 1 & 4 & 1 & 1/2 \\ 1 & 1 & 2 & 4 & 1 & 1/2 \\ 1 & 1/2 & 1 & 5 & 3 & 1/2 \\ 1/4 & 1/4 & 1/5 & 1 & 1/3 & 1/3 \\ 1 & 1 & 1/3 & 3 & 1 & 1 \\ 2 & 2 & 2 & 3 & 1 & 1 \end{bmatrix}$$

After Matlab calculations weight of the secondary indicator is given by $\begin{bmatrix} 0.25 & 0.75 \end{bmatrix}$. The weight of the three-level indicator at the strategic level is given by [0.4910, 0.0706, 0.2929, 0.1454], where CI = 0.0846, CR = CI/RI = 0.0846/0.90 = 0.0940 < 0.1, Consistency check has been done and it passed.

The weight of the three-level indicators at the tactical level is given by [0.1584, 0.1892, 0.1908, 0.0483, 0.1502, 0.2558], where CI = 0.0841, CR = CT/RI = 0.0481/1.24 = 0.0678 < 0.1, Consistency check have been done it also passed. The weights of the indicators are computation are shown in Table 5.

Table 5 Three-level indicator weights

Evaluation target index	Second-level index	Weight	Third-level	Weight
Lemon Company's cross-border e-commerce logistics distribution model selection	Strategic	0.25	National policy	0.4910
			Export tax laws and regulations	0.0706
			Consumer environment	0.2929
			Logistics technology	0.1454
	Tactical	0.75	Product properties	0.1584
			The price of logistics	0.1892
			Timeliness of logistics	0.1908
			Service quality of logistics	0.0483
			Enterprise strength	0.1502
			Seller demand	0.2558

3.2 The Relevant Gray Evaluation Numbers Are Calculated as Follows

The number of gray evaluations for the international express delivery model is as follows [2, 4–8]:

1. In the first gray level of this mode [8]: Pass the expert's score on international express through the first gray function [7].

$$\varphi 1\left(\mathrm{f}_{ijr}\right) = \begin{cases} 1 & f_{ijr} \geq 9 \\ \frac{f_{ijr}}{9} & 0 < f_{ijr} < 9 \\ 0 & f_{ijr} \leq 0 \end{cases}$$

   And gray rating function:

$$X_{ijr} = \varphi x\left(f_{ij1}\right) + \varphi x\left(f_{ij2}\right) + \cdots + \varphi x\left(f_{ijm}\right)$$

   where from analysis we got $X_{111} = 34/9$, $X_{121} = 37/9$, $X_{131} = 30/9$, $X_{141} = 38/9$, $X_{211} = 25/9$, $X_{221} = 32/9$, $X_{231} = 24/9$, $X_{241} = 19/9$, $X_{251} = 26/9$, $X_{261} = 28/9$.

2. By the second grayscale function:

$$\varphi 2\left(\mathrm{f}_{ijr}\right) = \begin{cases} \frac{f_{ijr}}{7} & 0 < f_{ijr} < 7 \\ \frac{14 - f_{ijr}}{9} & 7 \leq f_{ijr} < 14 \\ 0 & f_{ijr} \leq 0 \end{cases}$$

And gray rating function

$$X_{ijr} = \varphi x\left(f_{ij1}\right) + \varphi x\left(f_{ij2}\right) + \cdots + \varphi x\left(f_{ijm}\right)$$

From analysis we got, $X_{112} = 32/7$, $X_{122} = 33/7$, $X_{132} = 30/7$, $X_{142} = 32/7$, $X_{212} = 25/7$, $X_{222} = 32/7$, $X_{232} = 24/7$, $X_{242} = 19/7$, $X_{252} = 26/7$, $X_{262} = 28/7$.

3. By the third gray function:

$$\varphi 3\left(\mathrm{f}_{ijr}\right) = \begin{cases} \frac{f_{ijr}}{5} & 0 < f_{ijr} < 5 \\ \frac{10 - f_{ijr}}{9} & 5 \le f_{ijr} < 10 \\ 0 & f_{ijr} \le 0 \end{cases}$$

And gray rating function

$$X_{ijr} = \varphi x\left(f_{ij1}\right) + \varphi x\left(f_{ij2}\right) + \cdots + \varphi x\left(f_{ijm}\right)$$

From analysis we got, $X_{113} = 16/5$, $X_{123} = 13/5$, $X_{133} = 20/5$, $X_{143} = 12/5$, $X_{213} = 21/5$, $X_{223} = 18/5$, $X_{233} = 22/5$, $X_{243} = 19/5$, $X_{253} = 22/5$, $X_{263} = 22/5$.

4. Through the fourth gray function:

$$\varphi 4\left(\mathrm{f}_{ijr}\right) = \begin{cases} \frac{f_{ijr}}{3} & 0 < f_{ijr} < 3 \\ \frac{6 - f_{ijr}}{9} & 3 \le f_{ijr} \le 6 \\ 0 & f_{ijr} \le 0 \text{ or } f_{ijr} > 6 \end{cases}$$

And gray rating function

$$X_{ijr} = \varphi x\left(f_{ij1}\right) + \varphi x\left(f_{ij2}\right) + \cdots + \varphi x\left(f_{ijm}\right)$$

From analysis we got, $X_{114} = 2$, $X_{124} = 0$, $X_{134} = 11/3$, $X_{144} = 0$, $X_{214} = 10/3$, $X_{224} = 3$, $X_{234} = 3$, $X_{244} = 10/3$, $X_{254} = 11/3$, $X_{264} = 3$.

5. Through the fifth gray function:

$$\varphi 5\left(\mathrm{f}_{ijr}\right) = \begin{cases} 1 & 0 < f_{ijr} < 1 \\ 2 - f_{ijr} & 1 \le f_{ijr} \le 2 \\ 0 & f_{ijr} \le 0 \text{ or } f_{ijr} > 2 \end{cases}$$

And gray rating function

$$X_{ijr} = \varphi x\left(f_{ij1}\right) + \varphi x\left(f_{ij2}\right) + \cdots + \varphi x\left(f_{ijm}\right)$$

From analysis we got, $X_{115} = 0$, $X_{125} = 0$, $X_{135} = 0$, $X_{145} = 0$, $X_{215} = 0$, $X_{225} = 0$, $X_{235} = 0$, $X_{245} = 0$, $X_{255} = 0$, $X_{265} = 0$.

6. Calculate the gray evaluation weight vector and weight matrix of the country's dike form are given by [2, 7, 8]

$$X_{11} = \sum_{\varepsilon=1}^{5} X_{11\varepsilon} = X_{111} + X_{112} + X_{113} + X_{114} + X_{115} = 13.549,$$

Similarly,

$$X_{12} = 11.425, X_{13} = 15.286, X_{14} = 11.196.$$

So the matrix R_1 is formed as:

$$R_1 = \begin{bmatrix} 0.2788 & 0.3374 & 0.2362 & 0.1476 & 0 \\ 0.3598 & 0.4126 & 0.2276 & 0 & 0 \\ 0.2181 & 0.2804 & 0.2617 & 0.2399 & 0 \\ 0.3771 & 0.4083 & 0.2144 & 0 & 0 \end{bmatrix}$$

Similarly,

$$X_{21} = 13.883, X_{22} = 14.727, X_{23} = 13.495, X_{24} = 11.959,$$
$$X_{25} = 14.670, X_{26} = 14.511.$$

So the matrix R_2 is formed as

$$R_2 = \begin{bmatrix} 0.2001 & 0.2572 & 0.3025 & 0.2401 & 0 \\ 0.2414 & 0.3104 & 0.2444 & 0.2037 & 0 \\ 0.1976 & 0.2541 & 0.3260 & 0.2223 & 0 \\ 0.1765 & 0.2270 & 0.3178 & 0.2787 & 0 \\ 0.1969 & 0.2532 & 0.2999 & 0.2499 & 0 \\ 0.2144 & 0.2757 & 0.3032 & 0.2067 & 0 \end{bmatrix}$$

Similarly, the calculation of the packet direct mail mode is as follows:

$$X_{1i} = \begin{bmatrix} 11.3175 & 11.6032 & 10.7301 & 12.0063 \end{bmatrix}^{T} (i = 1, 2, 3, 4)$$
$$X_{2j} = \begin{bmatrix} 12.4126 & 12.2921 & 11.3714 & 12.6540 & 12.1714 & 12.5651 \end{bmatrix}^{T} (j = 1, 2, 3, 4, 5, 6)$$

$$R_1 = \begin{bmatrix} 0.3436 & 0.3913 & 0.2651 & 0 & 0 \\ 0.3352 & 0.4063 & 0.2584 & 0 & 0 \\ 0.4126 & 0.3994 & 0.1864 & 0 & 0 \\ 0.2869 & 0.3689 & 0.3165 & 0.0278 & 0 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 0.1790 & 0.2302 & 0.3223 & 0.2685 & 0 \\ 0.1717 & 0.2208 & 0.3091 & 0.2983 & 0 \\ 0.3518 & 0.4020 & 0.2462 & 0 & 0 \\ 0.1932 & 0.2484 & 0.3477 & 0.2107 & 0 \\ 0.1643 & 0.2113 & 0.2958 & 0.3286 & 0 \\ 0.2564 & 0.3297 & 0.3343 & 0.0796 & 0 \end{bmatrix}$$

Calculating overseas storage:

$$X_{1i} = \begin{bmatrix} 11.1936 & 11.9523 & 11.1963 & 11.8666 \end{bmatrix}^T (i = 1, 2, 3, 4)$$

$$X_{2j} = \begin{bmatrix} 11.7746 & 12.3365 & 11.3714 & 12.4953 & 12.2540 & 12.5333 \end{bmatrix}^T (j = 1, 2, 3, 4, 5, 6)$$

$$R_1 = \begin{bmatrix} 0.3772 & 0.4084 & 0.2144 & 0 & 0 \\ 0.2789 & 0.3586 & 0.3347 & 0.0279 & 0 \\ 0.3579 & 0.4083 & 0.2144 & 0 & 0 \\ 0.1966 & 0.1966 & 0.3539 & 0.2528 & 0 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 0.3020 & 0.3640 & 0.3057 & 0.0283 & 0 \\ 0.2342 & 0.3011 & 0.3567 & 0.1081 & 0 \\ 0.3518 & 0.4020 & 0.2462 & 0 & 0 \\ 0.2134 & 0.2744 & 0.3521 & 0.1601 & 0 \\ 0.1995 & 0.2565 & 0.3264 & 0.2176 & 0 \\ 0.1862 & 0.2394 & 0.3351 & 0.2394 & 0 \end{bmatrix}$$

The gray category calculation for international express delivery indicators is as follows:

$$B_1 = \omega_1 \times R_1 = [0.4910,\ 0.0706,\ 0.2929,\ 0.1454]^T \times \begin{bmatrix} 0.2788 & 0.3374 & 0.2362 & 0.1476 & 0 \\ 0.3598 & 0.4126 & 0.2276 & 0 & 0 \\ 0.2181 & 0.2804 & 0.2617 & 0.2399 & 0 \\ 0.3771 & 0.4083 & 0.2144 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.2810 & 0.3363 & 0.2399 & 0.1427 & 0 \end{bmatrix}$$

$$B_2 = \omega_2 \times R_2 = [0.1584,\ 0.1892,\ 0.1908,\ 0.0483,\ 0.1502,\ 0.2558]^T \times \begin{bmatrix} 0.2001 & 0.2572 & 0.3025 & 0.2401 & 0 \\ 0.2414 & 0.3104 & 0.2444 & 0.2037 & 0 \\ 0.1976 & 0.2541 & 0.3260 & 0.2223 & 0 \\ 0.1765 & 0.2270 & 0.3178 & 0.2787 & 0 \\ 0.1969 & 0.2532 & 0.2999 & 0.2499 & 0 \\ 0.2144 & 0.2757 & 0.3032 & 0.2067 & 0 \end{bmatrix}$$

$$= \left[0.2080\ 0.2675\ 0.2943\ 0.2229\ 0\right]$$

The calculation of the secondary indicators for direct mail parcels is as follows:

$$\begin{aligned}
\mathrm{B}_1 &= \omega_1 \times R_1 = [0.4910,\ 0.0706,\ 0.2929,\ 0.1454]^{\mathrm{T}} \\
&\times \begin{bmatrix} 0.3436 & 0.3913 & 0.2651 & 0 & 0 \\ 0.3352 & 0.4063 & 0.2584 & 0 & 0 \\ 0.4126 & 0.3994 & 0.1864 & 0 & 0 \\ 0.2869 & 0.3689 & 0.3165 & 0.0278 & 0 \end{bmatrix} \\
&= \left[0.3549\ 0.3914\ 0.2490\ 0.0040\ 0\right]
\end{aligned}$$

$$\begin{aligned}
\mathrm{B}_2 &= \omega_2 \times R_2 = [0.1584,\ 0.1892,\ 0.1908,\ 0.0483,\ 0.1502,\ 0.2558]^{\mathrm{T}} \\
&\times \begin{bmatrix} 0.1790 & 0.2302 & 0.3223 & 0.2685 & 0 \\ 0.1717 & 0.2208 & 0.3091 & 0.2983 & 0 \\ 0.3518 & 0.4020 & 0.2462 & 0 & 0 \\ 0.1932 & 0.2484 & 0.3477 & 0.2107 & 0 \\ 0.1643 & 0.2113 & 0.2958 & 0.3286 & 0 \\ 0.2564 & 0.3297 & 0.3343 & 0.0796 & 0 \end{bmatrix} \\
&= \left[0.2276\ 0.283\ 0\ 0.3032\ 0.1789\ 0\right]
\end{aligned}$$

The overseas storage calculation is as follows:

$$\begin{aligned}
\mathrm{B}_1 &= \omega_1\ \times R_1\ = [0.4910,\ 0.0706,\ 0.2929,\ 0.1454\ ]^{\mathrm{T}} \\
&\times \begin{bmatrix} 0.3772 & 0.4084 & 0.2144 & 0 & 0 \\ 0.2789 & 0.3586 & 0.3347 & 0.0279 & 0 \\ 0.3579 & 0.4083 & 0.2144 & 0 & 0 \\ 0.1966 & 0.1966 & 0.3539 & 0.2528 & 0 \end{bmatrix} \\
&= \left[0.3383\ 0.3740\ 0.2432\ 0.0387\ 0\right]
\end{aligned}$$

$$\begin{aligned}
\mathrm{B}_2 &= \omega_2\ \times R_2 = [0.1584,\ 0.1892,\ 0.1908,\ 0.0483,\ 0.1502,\ 0.2558\ ]^{T} \\
&\times \begin{bmatrix} 0.3020 & 0.3640 & 0.3057 & 0.0283 & 0 \\ 0.2342 & 0.3011 & 0.3567 & 0.1081 & 0 \\ 0.3518 & 0.4020 & 0.2462 & 0 & 0 \\ 0.2134 & 0.2744 & 0.3521 & 0.1601 & 0 \\ 0.1995 & 0.2565 & 0.3264 & 0.2176 & 0 \\ 0.1862 & 0.2394 & 0.3351 & 0.2394 & 0 \end{bmatrix} \\
&= \left[0.2472\ 0.3043\ 0.3146\ 0.1266\ 0\right]
\end{aligned}$$

International Express Score Vector is

$$\mathrm{H}_1 = \omega \times R = \begin{bmatrix} 0.25 \; 0.75 \end{bmatrix} \times \begin{bmatrix} 0.2810 \; 0.3363 \; 0.2399 \; 0.1427 \; 0 \\ 0.2080 \; 0.2675 \; 0.2943 \; 0.2229 \; 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0.2263 \; 0.2847 \; 0.2807 \; 0.2028 \; 0 \end{bmatrix}$$

Parcel direct mail score vector is

$$\mathrm{H}_2 = \omega \times R = \begin{bmatrix} 0.25 \; 0.75 \end{bmatrix} \times \begin{bmatrix} 0.3549 \; 0.3914 \; 0.2490 \; 0.0040 \; 0 \\ 0.2276 \; 0.2830 \; 0.3032 \; 0.1789 \; 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0.2594 \; 0.3101 \; 0.2897 \; 0.1352 \; 0 \end{bmatrix}$$

Overseas storage score vector is

$$\mathrm{H}_3 = \omega \times R = \begin{bmatrix} 0.25 \; 0.75 \end{bmatrix} \times \begin{bmatrix} 0.3383 \; 0.3740 \; 0.2432 \; 0.0387 \; 0 \\ 0.2472 \; 0.3043 \; 0.3146 \; 0.1266 \; 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0.2700 \; 0.3217 \; 0.2968 \; 0.1046 \; 0 \end{bmatrix}$$

Convert rating values to vectors C, C = [9, 7, 5, 3, 1].

AHP process designed by Satty and Vargas [9] have been followed for analytical hierarchy. It has been found the weight value of the overseas warehousing and distribution model (Z_1), parcel direct mail (Z_2) and international express delivery (Z_3) are given by:

$$\mathrm{Z}_1 = \mathrm{H}_1\mathrm{C}^{\mathrm{T}} = \begin{bmatrix} 0.2594 \; 0.3101 \; 0.2897 \; 0.1352 \; 0 \end{bmatrix} \times \begin{bmatrix} 9 \; 7 \; 5 \; 3 \; 1 \end{bmatrix}^{\mathrm{T}} = 6.0451$$
$$\mathrm{Z}_2 = \mathrm{H}_2\mathrm{C}^{\mathrm{T}} = \begin{bmatrix} 0.2263 \; 0.2847 \; 0.2807 \; 0.2028 \; 0 \end{bmatrix} \times \begin{bmatrix} 9 \; 7 \; 5 \; 3 \; 1 \end{bmatrix}^{\mathrm{T}} = 6.3594$$
$$\mathrm{Z}_3 = \mathrm{H}_3\mathrm{C}^{\mathrm{T}} = \begin{bmatrix} 0.2700 \; 0.3217 \; 0.2968 \; 0.1046 \; 0 \end{bmatrix} \times \begin{bmatrix} 9 \; 7 \; 5 \; 3 \; 1 \end{bmatrix}^{\mathrm{T}} = 6.4797$$

4 Result and Discussion

According to the above results, Z1 = 6.0451, Z2 = 6.3594, and Z3 = 6.4797. Where Z1 < Z2 < Z3. It can be concluded that the overseas warehousing and distribution model has the highest comprehensive score, followed by parcel direct mail and finally international express delivery. The results are related to the selection of the Lemon Company's cross-border export e-commerce logistics distribution model.

Based on the previous indexing system construction and the results obtained by using the analytic hierarchy process and gray correlation method, a comprehensive evaluation of the current logistics and distribution model existing in the cross-border e-commerce of Lemon Company have been done at strategic and tactical level. Among the cross-border e-commerce logistics distribution models, the company prefers overseas warehousing and distribution models.

Overseas warehousing is characterized by fast turnover and one-stop sorting, packaging, and distribution functions, which also meets the company's current main export product needs. Therefore, in terms of the long-term development of Lemon Company, it can increase its choice of overseas warehousing and distribution models. The development of Lemon Company's overseas warehousing model can also solve the existing problems of Lemon Company's logistics:

4.1 Ontime Distribution of Logistics

The overseas warehousing distribution mode can greatly reduce the logistical delivery time of Lemon Company. At present, most of Lemon Company's distribution mode is completed by international postal parcels, and it takes around 13–15 days. With the overseas storage mode, the goods can be delivered to customers in a maximum of 3 working days, which can improve the customer's satisfaction to Lemon Company's products to a certain extent.

4.2 Solve the Problem of Return

The overseas warehousing and distribution model can solve the problem of customers wanting to return products due to quality and other issues. Its existence would allow customers to return the goods directly to overseas warehouses, without the need of a series of complicated procedures such as customs clearance. It will also reduce the customer's burden on return costs which indirectly improves the service quality of Lemon Company. In the distribution process, the number of turnovers of the goods is also reduced, thereby reducing the damage rate of the goods packaging and the like, which can reduce the return rate of the goods.

4.3 Reduce Logistics Costs

Some of the fast selling goods such as some towels, soaps Can be stored in overseas warehouses and distribute it in batches, so that its average logistics cost will be reduced, hence the same product can be prevented from being sent frequently. Customs clearance and other steps can also be avoided. At present, Lemon Company uses the international express logistics distribution mode for the delivering the fast-selling products. In this method the logistics cost of small quantity will be higher than its own value.

5 Conclusion

Lemon's Company existing cross-border e-commerce logistics distribution model, overseas storage, international postal parcels and international courier, relies too much on direct postal parcel and direct mail which limits the development of Lemon company to a certain extent. Through model construction and analysis result, it shows that this company will be benefited through overseas storage model. This type of model can also be an example for other such companies. With the rapid development of cross-border e-commerce, more logistics distribution modes may appear. Lemon may need to recalculate the choice when selecting the cross-border e-commerce logistics distribution mode or may even look into advanced technological based logistics solutions such as those using developments like Blockchain [10].

References

1. B. Chen, Empirical study on cross border e-commerce enterprise logistics model under the background of economic globalization. Rev. Fac. Ing. **32**(12) (2017)
2. A. Yadav, M. Ali, M. Aris, S. Tuladhar, The analytical hierarchy process (AHP) approach for assessment of tax and revenue for Nepal. Int. J. Curr. Res. (IJCR) **7**(5), 15891–15896 (2015)
3. F. Ji, X.H. Zang, Innovation and development trend of cross-border e-commerce logistics. China Bus. Market**29**(6) (2015)
4. A. Yadav, M. Anis, M. Ali, S. Tuladhar, Analytical hierarchy process (AHP) for analysis: selection of passenger airlines for Gulf country. Int. J. Sci. Eng. Res. **6**(3), 379–389 (2015)
5. A. Yadav, G. Bhandari, D. Ergu, M. Ali, M. Anis, Supplier selection by AHP in KMC pharmaceutical: use of GMIBM method for inconsistency adjustment. J. Manage. Res. **7**(5), 19–46 (2015)
6. M. Ali, A. Yadav, M. Anis, Assessment of hazardous waste management proposal: using the analytic hierarchy process. Int. J. Econ. Comm. Manage. **3**(7), 315–327 (2015)
7. J. Pérez, Some comments on Saaty's AHP. Manage. Sci. **41**(6), 1091–1095 (1995)
8. Y. Su, Y. Wang, C. Mim, The forecast of development prospects of China's cross-border E-commerce based on grey system theory, in *2017 International Conference on Grey Systems and Intelligent Services (GSIS)* (IEEE, 2017, August), pp. 182–186
9. T.L. Saaty, L.G. Vargas, *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*, vol 175 (Springer Science & Business Media, Berlin, 2012)
10. R.R. Vokerla, B. Shanmugam, S. Azam, A. Karim, F.D. Boer, M. Jonkman, F. Faisal, An overview of blockchain applications and attacks, in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)* (2019)

The Construction of “Two-Side” Value Chain Between the Power Grid Side and the User Side in the Internet Trading Platform with Positive Sum Game Nature

Xin Yin and Shaozhen Hong

Abstract Seeking collaborative development between the power grid side and the user side can improve the operational capability of the power system. The theoretical and practical significance of this paper is to promote the development of management informatization of power grid industry and reduce the cost expenditure of power grid enterprises, as well as reduce the power loss. Based on the relationship between the power grid side and user side, we put up “the power grid side—Internet platform—the users side” with the game value chain. At the same time, we build up five value chain implementation plans: firstly, construct a multidimensional accurate portraits user model of power system; Secondly, carry out low pressure area weight overload warning of Internet users in the modeling; Thirdly, try to design and develop internet-based resource scheduling tools for user demand response; Fourthly, carry out invitation mechanism and dynamic adjustment; Lastly, design personalized rebate incentive strategies. Moreover, the positive-sum game relation and realization mechanism are considered, and the feasibility of value chain construction is described in detail with the value chain flow chart and technical flow chart.

Keywords The value chain · Power grid side and user side · The internet trading platform · Positive sum game

1 Introduction

With the rapid development of the economy and the increasing of the gross national product (GNP), The peak-valley difference in power system gradually increases and the peak load duration is short. Factors such as emergency, seasonal replacement, holiday factors, production and life rules, load composition affect the load peak and

X. Yin
Grid Chongqing Electric Power Co. Electric Power Research Institute, Chongqing, China
e-mail: 519759787@qq.com

S. Hong (✉)
School of Economics and Management, Beijing Institute of Graphic Communication, Beijing, China
e-mail: 15717260757@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_43

valley difference value of the power system. In 2020, according to the data released by the national energy administration, the domestic electricity consumption of urban and rural residents in China was 98.8 billion KWH in February this year due to the impact of COVID 19, which was up 3.1% year-on-year. In addition, there are great differences in the impact of power consumption in different provinces. Among them, the growth of power consumption in northern China was 12.3%, that in northeast China was 10.6%, and that in northwest China was 9.5%. The difference in peak and valley load of power system between provinces was between 30 and 40% of the maximum load [1]. With the increase of domestic electricity consumption caused by the epidemic, domestic electricity consumption increased rapidly. Therefore, it is an effective measure to explore a "two-sided combat" energy consumption reduction model between the power grid side and the user side, which serve the economic and social development.

2 Review of Related Studies

On the question of power grid side and user side, some scholars from the perspective of the peak load shifting that electrochemical energy storage system applied in peak power demand side cut scenes. For power grid side, the benefit is mainly to include slowing down power grid construction and reducing line loss, etc., the user side mainly is through the calculation of peak valley price straight after profits [2]. From the perspective of the user side, Guo Li et al. proposed that the time-of-use electricity price could improve the peak-valley difference rate by more than 15% and effectively reduce the negative power grid by studying the time-of-use electricity price and the optimal dispatching of electric vehicles, which is good for Load fluctuations [3]. From the perspective of the *State Grid*, state grid has also responded to the call to resolutely implement the decisions and arrangements of the CPC central committee and the state council. They studied and introduced eight measures and implemented the policy of phased reduction of electricity costs, resulting in reducing and exempting the electricity bill by about 48.9 billion yuan [4]. Also some scholars put forward a series of models. The current research on model focused on peak load shifting system. In order to maximize the user's earnings, Han Chao [5] introduced the model of battery energy storage system and put forward a set of optimization strategy on the existing electric power market system framework. Through adjusting the charge system users to maximize the benefits at the same time, this framework reached a certain load shifting effect [5]. However, this paper focuses on the maximization of user benefits, and does not establish a clear model for solving the peak and filling problem. Yong [6] established the energy storage system which was behaved as the energy cost minimum, and peak load mathematical model of the cost of the small target, the determination of the specific charge and discharge strategy, so as to achieve peak load shifting effect. But what the literature mainly considered is using energy storage system adjusted by the cold and hot load of the building, and it did not consider building peak power load system of load shifting process [6].

As you can see, at present, research on power grid side and user side, mostly concentrated in the “grid active type” category. It is relatively small about guiding the client, digging the subjective initiative of the related research and the research on the user side peak load shifting products or measures focused on the user side peak parameter and load shifting energy storage device. The peak valley price guides users to participate in peak load shifting, and based on electric power demand side management platform market is proposed, but this paper is only based on the Internet-based trading platform to realize power grid peak load shifting, thinking little about energy conservation and consumption reduction research.

In this paper, an economic theory based on a positive-sum game is constructed, and a value chain on the power grid side and the user side is constructed based on the quasi-real-time transaction model of The Internet-based trading platform. Positive-sum game refers to a state in which the two-sides of the game achieves a “win–win” situation in the market competition by cooperating with each other. A situation where the sum of the gains is positive [7]. And the concept of the value chain was first established in 1985 by Michael porter’s *Competitive Advantages* [8]. Here, we refer to the circumstance of power grid industry, combining with the porter put forward the value chain theory as the power grid side and user side as a whole, and we know that both economic and social ties are inseparable which fully embodies the value chain. Therefore, combined with game theory and value chain, we might as well put the power grid side and user side understood as quasi real-time scheduling and intelligent optimization of electric power resources configuration. We can put the user side subjective active power saving effect in energy saving and protection of category into consideration thus active energy and electric potential can be received. From the dimension of “Different but interrelated energy consumption activities”. We establish a value creation based on power grid side and user side of the dynamic process in the structure, which can be namely “double side” with the game value chain.

3 Construction of “Double Side” Positive Sum Game Value Chain

3.1 Construction of “Double Side” Positive Sum Game Value Chain

Considering the above research, this article is based on the premise of realizing accurate load forecasting. From establishing the Internet-based trading platform ideas of quasi real-time transaction model and according to the electricity information collection system of “96 points” user side to collect information on time electricity, power grid enterprises relied on WeChat, Alipay, online state grid, electric E, palm on the power to realize “Internet platform side of peak shaving, electricity peak, reducing the load peak” with game mechanism. In the process of achieving work, power grid enterprises tend to pursuit of self-interest maximization. There will be toward to the direction of their own interests game. Therefore, to enhance the interests of the power grid side and the user side and the whole society, the "double side" on the value chain must establish the effective mechanism, which can suppress the related interest parties in electricity use excessive and promote their effective cooperation, making the game's final effect tend to be the "positive sum game", and then the related parties combination achieves the optimal strategy of *Nash Equilibrium.*

On the value chain of "power grid side—Internet platform—user side", the specific ideas are as follows:

- Promote peak-load reduction and valley filling of the power grid and reduce the loss reduction benefits generated by energy losing of power equipment. Based on the Internet-based trading platform and the scientific guidance and change of users' electricity behavior, the power grid side promotes the power grid to shift peak and fill valley and it improves the utilization rate of equipment, reducing line loss. In other words, the Internet platform is used to realize the "complete information market of power grid" between the power grid side and the user side, so that the user side can fully understand the power strategy space and strategy combination of the power grid side, and then promote the realization of positive sum game.
- Avoidable capacity are benefits generated by promoting the grid load curve to be flat and reducing or delaying the grid investment. By means of peak load shifting, the load curve of the power grid can be flattened, and the time-peak-valley difference can be reduced, so as to delay or reduce the huge investment of passively increasing the equipment capacity and building new substation due to the non-electric power shortage in the short-term peak period. In other words, the power grid side and the user side can participate in the power management at the same time, so as to realize the balanced strategy beneficial to the power utilization on the load of the power grid, and effectively avoid the conditional strategic actions of both sides.
- Alleviate the peak load and supply pressure of the power grid and reduce the avoidable benefit of purchasing peak power. The innovative design of "peak avoidance" powerful Internet trading products and reducing the peak load of electricity can reduce the power transaction costs to ease the peak period of power supply constraints. That is to say that avoiding the conditional strategy caused by the cost effective lead to the phenomenon of zero-sum game.

3.2 *“Two-Sided” Positive Sum Game Value Chain Construction Measures*

3.2.1 User Side: Construct a Multidimensional Accurate User Portrait Model of Power System

Using key technologies such as data integration, data cleaning, data verification, user feature extraction, multi-heterogeneous data fusion, we decide how to combine user registration information, Internet transaction information and historical telecommunications Information with use of big data technology to analyze users’ electricity behavior and preferences, for users to generate multidimensional precision portrait, while categorizing users in the resource pool.

3.2.2 Power Grid Side: Carry Out Early Warning Modeling of Overload Monitor in Low-Voltage Platform Area

Combining area of historical operation data, the area users and natural and social environment data with weight overload event occurrence time, duration, and degree of weight overload characteristic, analysis method is adopted to establish the area weight overload related events, select and extract the area weight overload characteristic through cluster analysis classifying weight overload event correlation; Machine learning technology is used to train the load overload prediction model, analyze and explore the key factors leading to the overload in the platform area, and finally we establish the load overload prediction model in the platform area, so as to discover the load overload area in advance, and it can help to form the closed-loop management of early warning, early prevention, in-process handling and post-analysis.

3.2.3 Internet Platform: Design and Develop Internet-Based Resource Scheduling Tools for User Demand Response

By using key technologies such as information management, information system development and computer network system integration, the paper designed and developed resource scheduling tools for resident users’ demand response based on an Internet platform. We select pilot programs for application demonstration and evaluate the implementation effect according to the demonstration results which can be carried out by extensive promotion.

3.2.4 Coordinated Development: Carry Out Precise Release of Invitation Mechanism and Intelligent Dynamic Adjustment

By using key technologies such as big data analysis and processing, self-learning of invitation mechanism based on artificial intelligence, and intelligent dynamic adjustment of invitation mechanism, we can combine with real-time peak adjustment demand of the power grid side. This paper proposes a dynamic multi-round invitation mechanism for the user side by adopting iterative thinking and artificial intelligence methods. Through iterative optimization, the rotation and time interval of peak avoidance is constantly improved.

3.2.5 Interest Driven: Design Multiple Personalized Rebate Incentive Strategies

The positive sum game emphasizes the symmetry of bilateral returns and the reinforcement incentive theory advocates more positive reinforcement and less negative reinforcement while maintaining the continuity of negative reinforcement [9]. Therefore, the power grid side can establish a differential rebate settlement model by using key technologies such as incentive response mechanism construction, diversified decision support model, multi-dimensional hierarchical intelligent incentive algorithm and so on, and we can comprehensively consider the user's contribution to the performance and response speed.

3.3 Construction of "Double Side" Positive Sum Game Value Chain

The positive-sum game belongs to the interest-driven behavior of both sides of the game. The game behavior in which the interests of one party increase while the interests of the other party are not damaged so as to realize the increase of the interests in the whole society. Most of the game behaviors discussed in economics based on the asymmetry of information. However, in the digital age, we can make use of the information openness of the Internet to construct the value chain of positive-sum game, so that the interests of the power grid and users can reach an equilibrium state.

3.3.1 Realization of "Two-Sided" Positive Sum Game

To implement the power grid side and user side with game, it is necessary to establish the corresponding mechanism and overcome which is likely to lead to z*ero game* result. The phenomenon is negative and the unfavorable for game development.

Table 1 List of positive sum game relations and realization of power grid side and user side

Game dimension	Energy of electrical equipment	Electric load	Peak power usage
Game focus	Whether to produce loss reduction benefit	Whether to produce avoidable capacity benefit	Whether to produce avoidable electricity benefit
Implementation mechanism	Checking system	Cooperative system	Incentive method
Main contents	Using Internet trading platforms to change electricity usage behavior	Participate in power management together	Reduce peak load
Measures	Power system user side portrait model	User side demand response resource scheduling	Heavy overload warning modeling in low voltage platform area Dynamic multi-round invitation mechanism on the user side Personalized rebate incentives
The ultimate goal	Equipment utilization is maximized and line loss is minimized	The load curve of the power grid is flat and the time difference between peak and valley is minimized	No power supply pressure, power transaction costs to the minimum
Risk measure	Both sides unilaterally pursue the interests of all parties	A collaboration barrier on both sides	A zero-sum game on both sides

Therefore, we combine with the survey data of *the State Grid Chongqing Electric Power Company Electric Power Research Institute* to make the following in same relation and the schedule, as showed in Table 1.

3.3.2 Framework Design

Taking the Internet platform as the "two-sided" connection hub, the following Fig. 1 is proposed based on the above measures.

3.3.3 Principle of Data Processing Process

We use big data processing system based on the Internet platform and user requirements with the characteristics of the data to realize the functions of classified query, data retrieval, data comparison, data correlation. And we analyze the two ends of the

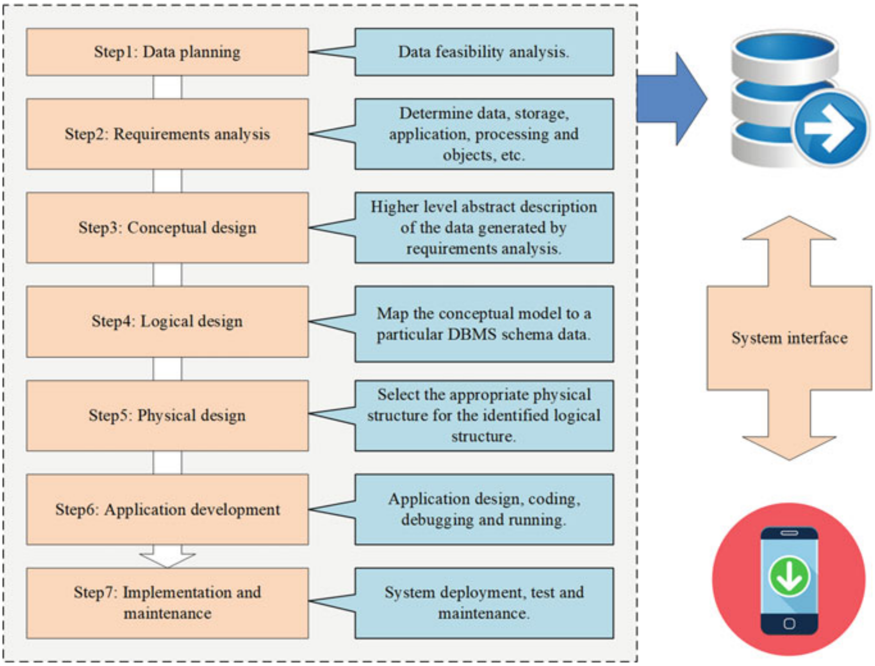

Step1: Data planning
Data feasibility analysis.
Step2: Requirements analysis
Determine data, storage, application, processing and objects, etc.
Step3: Conceptual design
Higher level abstract description of the data generated by requirements analysis.
Step4: Logical design
Map the conceptual model to a particular DBMS schema data.
Step5: Physical design
Select the appropriate physical structure for the identified logical structure.
Step6: Application development
Application design, coding, debugging and running.
Step7: Implementation and maintenance
System deployment, test and maintenance.
System interface

Fig. 1 Two-sided positive sum game value chain flow frame diagram

value chain with data, in order to specific implementation for the seven steps: data across gauge, demand analysis, concept design, logic design and physical design, application development, implementation and maintenance. Then the power grid side carries out user analysis after identity authentication to access control, data management and import with export, system data maintenance operations on the use of system realization Alipay, WeChat public number, WeChat applet articulated user side, so as to achieve the purpose of seamless power grid side and user side. By doing this, we can enrich the data collection, transmission and interactive methods for a system which is designed a unified mobile terminal new interface and interaction. As showing in Fig. 2.

4 Conclusion

The Internet-based trading platform trading pattern of integration of power grid side and user side on time with the game value chain is to develop social benefits and economic benefits, Power grid enterprise business communication, the Internet industry and online interoperability means of information technology. This method is currently the only way to realize digital power grid industry. It can also realize the power grid side and user side to get self-optimization, controlling with virtual to real,

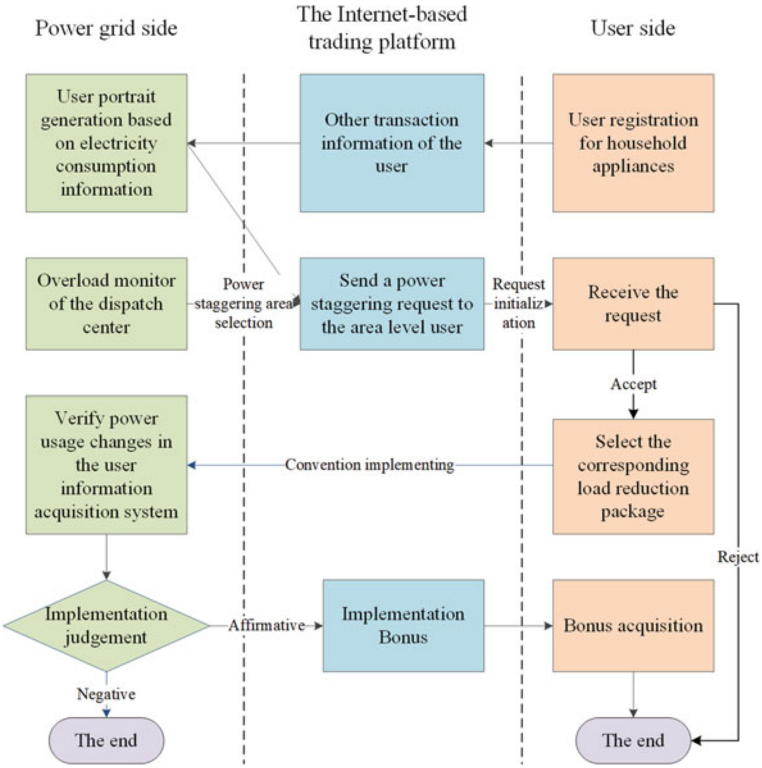

Power grid side
The Internet-based trading platform
User side
User portrait generation based on electricity consumption information
Other transaction information of the user
User registration for household appliances
Overload monitor of the dispatch center
Power staggering area selection
Send a power staggering request to the area level user
Request initializ ation
Receive the request
Accept
Verify power usage changes in the user information acquisition system
Convention implementing
Select the corresponding load reduction package
Reject
Implementation judgement
Affirmative
Implementation Bonus
Bonus acquisition
Negative
The end
The end

Fig. 2 Technical flow chart of Internet path data processing

mixing for decision-making and implementation as a way of intelligent revolution. The author believes that, with the continuous improvement of user network scale, the power grid side will bring the positive-sum game with alleviating the peak pressure of power grid, improving equipment utilization and insuring safe and reliable power supply, reducing loss, delaying of grid construction investment benefit at the same time. The construction of the value chain will inevitably promote the leap type development of informatization in the management of the power grid industry.

References

1. The growth rate of domestic electricity consumption will pick up obviously in March [EB/OL]. https://www.nea.gov.cn/2020-03/30/c_138931213.HTM.People.com.cn. 31 Mar 2020
2. L. Guo, G. Xue, C. Wu, Z. Xie, G. Liu, B. Li, Economic benefit analysis of energy storage system applied in peak load valley load shifting. Power Demand Side Manag. **21**(05), 31–34 (2019)
3. G. Yang, H. Luo, D. Wang, G. Fu, R. Gia, L. Yao, Master-slave game model of time-of-use electricity price and optimal dispatching of electric vehicles. J. Power Syst. Autom. **30**(10),

55–60 (2018)

4. X. Tian, State grid pays the national policy dividend to the actual situation. State Grid News (002) (2020)
5. C. Han, *Research on Optimal Scheduling of Energy Storage System for User Oriented Peak-clipping and Valley Filling [D].* (Zhejiang University, 2017).
6. Y. Sun, S. Wang, F. Xiao, D. Gao, Peak load shift control using different cold thermal energy storage facilities in commercial buildings: a review. Energy Convers. Manag. 71 (2013)
7. R. Liu, H. Wang, Application of game theory in market competition—a case study of zero-sum, negative sum and positive sum game. Gansu Sci. Technol. **48**(07), 63–66 (2019)
8. M.E. Porter, *The Competitive Advantage: Creating and Sustaining Superior Performance* (The C FreePress, New York, 1985)
9. Y. Ye, Research on game rule optimization from the perspective of ecological niche. Econ. Forum **05**, 131–134 (2016)

Mining and Analysis of Emergency Information on Social Media

Dan Chang, Lizhu Cui, and Yiming Sun

Abstract With the advent of the social media era, various social networking sites and social apps are growing at a high speed. As an important product of the WEB 2.0 era, Sina microblog has become an important vehicle for the dissemination of emergency information. In this paper, the textual features of microblogging are first analyzed and then text pre-processed based on the emergency response information of the microblog platform. Based on this, an MB-LDA (MicroBlog-Latent Dirichlet Allocation) topic model based on the "User-Document-Topic-Word" structure is proposed. The aim is to improve the government's ability to respond to emergencies and to improve the efficiency of government emergency information collection by thematically mining and analyzing emergency information in case of emergencies, so as to obtain the actual situation of emergencies and other effective emergency information.

Keywords Topic mining · Emergency information · Microblog · MB-LDA model

1 Introduction

In an era of rapid growth of the Internet and its increasing popularity, social media has come of age. In 2006, Twitter, the world's first micro-blogging platform, was born. After much researches and developments, Sina microblog came out in 2009. As an important product of the WEB2.0 era, microblog has developed rapidly. According to the data, as of the end of 2019, the monthly active users of Microblog reached 516 million, compared to the end of 2018, a net increase of about 54 million. The text

D. Chang · L. Cui (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 18120609@bjtu.edu.cn

D. Chang
e-mail: dchang@bjtu.edu.cn

Y. Sun
The Faculty of Engineering, The University of Melbourne, Melbourne, Australia
e-mail: yisun3@student.unimelb.edu.au

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_44

data on microblog has the characteristics of short text, and most of the microblog texts are the thoughts and feelings of individual users, and their terms are more colloquial and casual. In most cases, if the user wants to know information about a specific event or thing, this can be achieved by keyword search. However, it is precisely because of the particularity of the microblog text that what users search for is very fragmented. Data mining techniques are necessary to have complete information about an event or something. Therefore, the study of data mining methods in conjunction with the microblog context is urgently needed.

At present, China is in an important historical stage of economic and social transformation, and various emergencies occur frequently. How to effectively respond to various emergencies has aroused widespread concern from all sectors of society. In the process of dealing with emergencies, the government's collection of information on emergencies runs through the entire process of emergency response, which is not only the decision maker's response to emergencies, but also the first step in taking emergency measures, and is the key to success or failure in dealing with emergencies. The huge amount of information carried on microblogs has become an important data source, containing a wealth of research and application value. Therefore, this paper proposes MB-LDA (Microblog-Latent Dirichlet Allocation) topic model based on the analysis of textual characteristics of microblogs to mine and analyze the emergency information on Sina microblog, to help decision-makers obtain valuable information from Sina microblog more effectively and improve the efficiency of emergency information collection.

2 Literature Review

Mining social media messages using thematic models is a current research hotspot. The direct application of the LDA topic model to the study of texts under social media was not satisfactory, so scholars began to improve the traditional LDA topic model according to the different characteristics of different social media. As Zhao et al. proposed the Twitter-LDA topic model, using the central idea of Author-Topic Model to integrate all the tweets of the same user into a single document, data mining of user-level and tweet-level topic information based on the addition of the background model has achieved some results, but there is still a lot of noise [1]. To this end, many research scholars and experts have begun to delve into ways to reduce noise in documents. For example, Tu et al. proposed an improved detection mechanism by adding one or more classification processes based on Bayesian classification algorithms before the text clustering module [2]. To mine potential topic information, Zhao et al. proposed a topic model based on the potential Dirichlet Allocation (LDA)-Hashtag-LDA, which can greatly improve the amount of tags and vocabulary in microblogs by co-modeling them [3]. Also the impact of tagging on potential topic generation improves the accuracy of data mining. Based on the traditional text-level mining, Q. Wen proposed a user-level mining model, the Authorship Model (ATM), which can achieve efficient accelerated sampling without reducing accuracy [4].

Since Microblog is one of the most popular social media platforms and contains a wealth of information, many scholars have focused their research on the analysis and mining of short microblog posts using topic models. For example, Wang et al. proposed a Multiattribute Latent Dirichlet Allocation (MA-LDA) topic model that incorporates the temporal and hashtag attributes of microblogs into the LDA model to mine hot microblog topics [5]. Zhi et al. proposed a method to optimize the overlapping completeness evaluation method for the identification of microblogging topics by combining the quantitative effects of LDA and TF-IDF [6]. Yu et al. proposed a Dirichlet polynomial hybrid (DMM) topic model based on the internal data extension of the user-LDA topic model with the potential feature vector representation (ULW-DMM) of words trained in an external corpus. Experimental results show that the ULW-DMM model can detect microblogging topics more effectively [7]. Thus, the mining and analysis of microblog textual information has a rich research base.

There is a lot of valuable information to be mined from the emergency information that is passed on social media. Sakaki T et al. combined the knowledge of disaster emergency response to detecting the occurrence of disaster events [8]. Crooks A et al. used Twitter as a sensor to understand the situation after a disaster [9]. Han analyzed the microblog text related to the 2018 Shouguang flood disaster in terms of space, time, and content. Based on the LDA model and RT algorithm, a topic classification model was established to identify the flood-related topics in the microblog text [10].

To sum up, since the day of the rise of social media, not only has social media become an important source of data collection, but social media itself has also become an indispensable object of research. Based on the current state of the research, it can be found that the LDA topic model is a major method to study Twitter tweet text as well as Sina Microblog text. Further research is needed on the topic of microblog text mining, and few scholars have researched emergency information mining. Therefore, based on the short and sparse text of Sina microblog, this paper introduces the user dimension and builds a four-tier MB-LDA model of "text-user-topic-word". The model can mine and analyze emergency information and can provide suggestions for the allocation and deployment of public emergency resources by the relevant departments.

3 Microblog Mining Analysis

3.1 Microblog Text Feature Analysis

Microblog provides a completely open service platform where billions of users can speak out freely, and a huge amount of information appears on the platform every day. However, because of the highly interactive and open nature of microblogs, microblog texts are different from traditional data [11], and it is challenging to analyze them. Among them, microblogging texts have the following main characteristics.

3.1.1 Timeliness

The majority of microblog users embody their current state of mind in their microblog text, so such information is highly communicative and newsworthy in real time, and this value will diminish over time [12]. Besides, when searching for a keyword on the microblog, the last published content about that keyword will be presented in the first article, which is also an indication of the timeliness of the microblog text [13].

3.1.2 Sparseness

Each microblog text must be under 140 characters in length, whereas traditional text data is longer. Therefore, the sparsity of the microblog text data is due to the word limit of the microblog text. The sparse microblog textual information contains very few keywords, so the process of topic mining poses some challenges.

3.1.3 Non-standard

The irregularity of the microblog text is first of all because it contains a variety of information elements, and because of the word limit, users sometimes use pictures, videos and other multimedia information to replace the content to express the content clearly in a limited number of words. Secondly, users can express their opinions as they like through microblog, and the colloquial situation is serious. Besides, a lot of new words are emerging, such as “Koi” and “Guan Xuan”. For these reasons, the difficulty of keyword recognition is greatly increased, and it will also affect the difficulty of mining to a certain extent.

3.1.4 Repeatability

An important function of microblogs is “retweeting”, where people can retweet content of interest to more users, and on the other hand, when commenting on other people’s microblogs, they can also retweet their comments as a new microblog by checking the retweet box. Each of these situations allows the same information to be presented on different user home pages.

3.2 Text Pre-processing Analysis

Text pre-processing is the first and most important step in text mining. The accuracy of text mining is often determined by text pre-processing, as pre-processed text is often mined using similar data mining models. For microblogs, effective microblog text data has the meaning of topic mining, invalid microblog text data often refers to

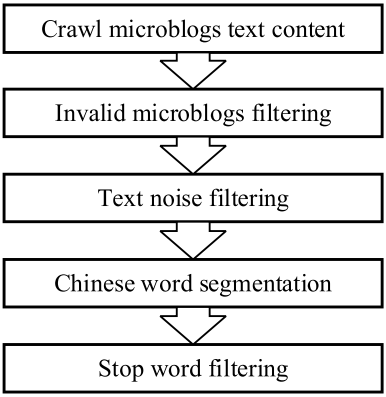

Crawl microblogs text content
Invalid microblogs filtering
Text noise filtering
Chinese word segmentation
Stop word filtering

Fig. 1 Main process of text pre-processing

the entire blog post only contains web links, images, color text [14]. Wherein, the main steps of text preprocessing are shown in Fig. 1.

3.2.1 Invalid Microblogs and Noise Filtering

The extremely open, real-time and interactive nature of the Microblog platform has led to the unregulated nature of the microblog text. In addition, some microblog users choose to share only pictures instead of posting substantive textual content, some users just post some colorful text to express their current state of mind, or some so-called "microbloggers" just post web links for marketing. These are considered invalid microblogging texts. The set of microblog texts after excluding invalid microblog texts in terms of formula (1) [15]:

$$\mathrm{D} = \{(d_i)|i = 1, 2, 3, \ldots, n\} \tag{1}$$

In this formula, the letter i represents the text number of the microblog, the letter d_i indicates that this is the i microblog, the letter n is the number of microblogs, and D represents the collection of all microblog texts.

Some non-standard microblog texts often do not contain topics, and whenever you want to notify a specific user in the blog post, the @ and other symbols will be used. These often do not have a positive effect on the experiment and will affect the accuracy of the experiment. It is called "noise". Therefore, to improve the mining efficiency of the model, these noise data must be filtered.

3.2.2 Chinese Word Segmentation Technology

Words are the smallest unit of text collection. The biggest difference between Chinese words and English words is that there is a space symbol in English words one by one to

Table 1 ICTCLAS Chinese part of speech tagging set

Part of speech	Part-of-speech tagging	Part of speech	Part-of-speech tagging
Adjective	a	Other proper nouns	nz
Distinguishing words	B	Onomatopoeia	o
Conjunction	C	Preposition	p
Adverb	d	Quantifier	q
Interjection	e	Pronoun	r
Morpheme words	g	Numeral	m
Predecessor	h	Common noun	n
Idiom	i	Position noun	nd
Abbreviation	j	Name	nh
Subsequent ingredients	k	Particle	u
Institution name	ni	Verb	v
Place noun	nl	Punctuation	wp
Place name	ns	String	ws
Time word	nt	Non-morpheme words	x

indicate the interval. There is no space symbol in Chinese to distinguish. At present, the Chinese word segmentation tool has begun to take shape, and the ICTCLAS Chinese word segmentation system is a Chinese word segmentation system that is used more frequently [16]. ICTCLAS is a word segmentation model based on the cascading hidden Markov principle. It can not only perform Chinese word segmentation but also perform many useful functions such as part-of-speech tagging and new word recognition. Among them, the ICTCLAS Chinese part-of-speech tag set is shown in Table 1.

The Chinese segmentation using the ICTCLAS Chinese segmentation system is referred to as a dictionary-based Chinese text segmentation. In addition to this dictionary-based Chinese text segmentation, there are two other methods of Chinese text segmentation, namely, statistical-based Chinese text segmentation and comprehension-based text segmentation.

The statistical-based Chinese word segmentation method, also known as the dictionary-free word segmentation method, calculates the co-presentation information of adjacent words in the corpus by calculating the co-presentation frequency of the adjacent words in the corpus. The word is considered a constitutive word when the interaction information exceeds a certain threshold. That is, the greater the probability of adjacent words coexisting, the greater the probability of forming words. However, this method can easily misjudge the phrases “my” and “one” as words, so the words need to be segmented in conjunction with the lexical method. According to the three basic principles of probability word frequency statistics, such as mutual information, multivariate statistical model and t-test principle, and using the Chinese word segmentation system designed by ourselves to compare. Combining dictionaries with statistical methods, first using the forward and backward maximum matching method for segmentation, and then processing the words according to the principle of combining statistics and rules can eliminate fuzzy fields, solve the problem of high-frequency unknown words, and solve combinations fuzzy fields of elimination. After preliminary segmentation of commonly used word frequency statistics data, the use of support vector machines to further segment fuzzy fields, and the use of mutual information to represent fuzzy fields can greatly improve the segmentation efficiency. Besides, the hidden Markov signal source can also be used to represent the signal source (that is, the binary statistical model) and perform rough segmentation to calculate the weight of the directed edges of the model [17, 18]. Then, the weight can be modified according to the length of the word. Finally, the shortest path method is used to obtain the segmentation result.

The word segmentation method based on comprehension, this method is to use the computer to simulate the sentence understanding process to identify words, while dividing the words, by analyzing the grammar and semantics, you can eliminate ambiguity. Semantic research is the key to theoretical and practical breakthroughs. The research of this method involves relevant knowledge such as artificial intelligence and expert system. This method of word segmentation has a good effect, but it is very complicated to implement. Learning and training require a lot of general and complex Chinese knowledge and information, which is difficult for the machine to read directly. Therefore, the development of this method is still immature and still in the experimental stage. At present, the main methods of artificial intelligence are expert system method, neural network method and generation test method. Among them, the word segmentation is based on the inference mechanism, and the heuristic knowledge base expressed by the production rules is used to cut the fuzzy field.

This paper uses the Chinese word segmentation method based on the dictionary because both the keywords of emergency events and the subject words of emergency information are the main verbs, adjectives and nouns. Therefore, when preprocessing the collected microblog text data, to improve the accuracy of mining, only verbs, adjectives and nouns are reserved.

3.2.3 Stopwords Filtering

In the text, not every word or character can represent the text, such as the English qualifier “the”, “this”, “those”, the Chinese modal words “ah”, “la”, “ma”, etc. They are not of any substance in themselves. They play only a secondary role in the text. They tend not to have any effect on the results of the text analysis and processing, then these words fall under the category of stopwords [19]. Stopwords often have no real meaning, but can appear in large numbers in Chinese texts. So, removing stopwords from the text can greatly increase the efficiency of late topic mining, which is a crucial step. The deletion of stopwords should follow two principles: one is to reduce the number of dimensions represented in the text by deleting meaningless stopwords. Second, the deleted stopwords do not affect the original semantics of the text [20]. Currently, the existing glossary of stopwords is not particularly effective, and the research results are relatively sparse.

The current stopwords lists include the Harbin Institute of Technology stopwords list, Sichuan University Machine Intelligence Laboratory stopwords list, and Baidu stopwords list. Among them, the Harbin Institute of Technology stopwords list contains 767 Chinese stopwords, the Sichuan University Machine Intelligence Laboratory stopwords list contains 976 Chinese stopwords, and the Baidu stopwords list contains Chinese and English stopwords, a total of 1395 stopwords [21]. There is no standard and comprehensive list of Chinese stopwords, and users need to process them according to their research subjects because some words can't be completely removed as stopwords [22–24].

Combining the characteristics of microblog text, and through the above analysis of stopwords, this paper defines the following rules in the process of filtering stopwords in microblog text:

1. Pronouns, auxiliary words, mood words, adverbs, and prepositions are filtered as stopwords.
2. The Pre-compositions, idioms, abbreviations, Post-compositions, and location words cannot be filtered as stopwords.
3. Nouns cannot be filtered as stopwords.
4. Retain the two types of symbols "@" and "#".
5. Remove all punctuation marks except rule d.

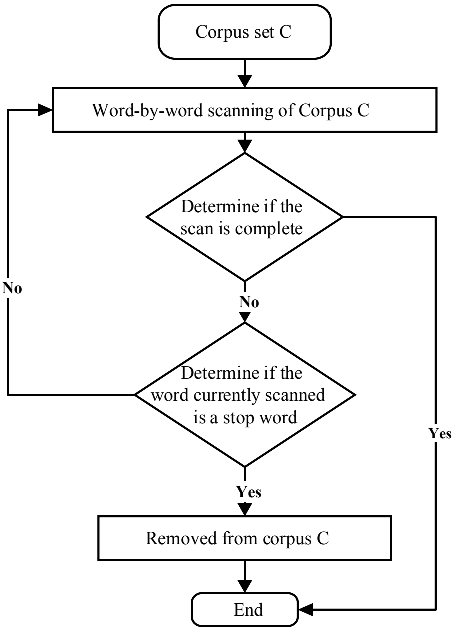
The specific process of its filtration is shown in Fig. 2.

3.3 Comparison of Microblog Topic Mining Methods

At present, the mainstream methods for microblog topic mining are the LDA topic model, the ATM model, and their improved various topic models. In his study, Hong compared and analyzed the topic results generated by the LDA topic model, ATM extension model, training-based user mode model, and term mode model [25]. Research has shown that the topics mined by different modeling approaches are not fully equivalent in terms of the consistency of the topics mined. Even if similar topics are obtained using different topic mining models, the probability of similar topics constituting words calculated using the Jensen-Shannon Divergence algorithm varies, and this variability increases with the number of topics detected. In terms of extended models, both the ATM model and the Twitter-LDA topic model are an extension of the traditional LDA topic model, and the Twitter-LDA topic model is another extension of the ATM model because the Twitter-LDA topic model can be modeled not only at the blog level but also at the user level.

The concept of "topics" and "super topics" has emerged in microblogs today, where users can click on the corresponding tags to view the content and post about

Fig. 2 Stop word filtering process


Corpus set C
Word-by-word scanning of Corpus C
Determine if the scan is complete
No
Determine if the word currently scanned is a stop word
No
Yes
Removed from corpus C
Yes
End

it for more users to know. The Labeled-LDA topic model is to introduce the concept of label information in the microblog. The Labeled-LDA topic model transforms the traditional unsupervised learning LDA topic model into a supervised learning model, which greatly improves the efficiency of topic mining modeling. Quercia et al. [26] also experimented and evaluated the Labeled-LDA topic model and found that when using the model to classify microblog users, they found that the text aggregation of posts from generally active users resulted in moderate document length, and Labeled-LDA had the best classification effect on such texts. Taken together, Table 2 has a more detailed description of the above-mentioned topic mining modeling methods.

4 Construction of MB-LDA Topic Model

4.1 Introduction of MB-LDA Topic Model

4.1.1 Model Construction Ideas

As you can see from the previous chapter, the text data of microblogging is more complex than the general text data, and the Microblog platform is also full of unstructured text data with unclear topics. Using the traditional LDA topic model for topic mining effect is not good, using Twitter-LDA topic model for topic mining will find that the model is mainly applied to Twitter text mining, and Twitter text is mainly

Table 2 Comparison of microblog topic mining models

Model name	Method to realize	Advantage	Limitation
LDA topic model	Use directly	No supervision	Mining effect is not ideal
ATM model	Text aggregation	Solve short text problems	Only user-level topic modeling
ATM extended model	Text aggregation	Solve short text problems	Mining topic are few and not ideal
Twitter-LDA model	Text gathering, introducing background model	Solve short text and high-frequency vocabulary problems	One post corresponds to one topic
Labeled-LDA model	Introduce label information	Improve topic interpretability	Must have sufficient label information

Table 3 Definition of relationship

Type	Symbol	Description
The contact relationship	@	User Relationship refers to the potential semantic relationship between microblogs with @ and contacts with @. In general, microblogs that are associated with the same contact are often related to their topic
The textual relationship	RT	Textual relationship refers to the potential semantic relationship between the microblog with RT and the original microblog. In general, the topics in the forwarding section and the original section tend to be related

English, and Sina Microblog blog posts are mainly Chinese, both in the language style and language structure are very different, so the use of Twitter-LDA model is not an effective choice. Therefore, the LDA topic model was improved by combining the features of Twitter-LDA and microblogging text, and the MB-LDA topic model was used to mine microblogging text.

Microblogs, unlike ordinary text, carry textual information that represents the relationship between microblogs: @ and RT. @ represent the user relationship between microblogs, while RT represents the textual relationship between microblogs, which are defined as follows (Table 3).

4.1.2 Construction of MB-LDA Topic Model

The traditional LDA topic model is a three-layer Bayesian probability model, and considering the user-level associations, MB-LDA is a four-layer Bayesian probability model, namely “user-document-topic-word”. The Bayesian network diagram of the model is shown in Fig. 3.

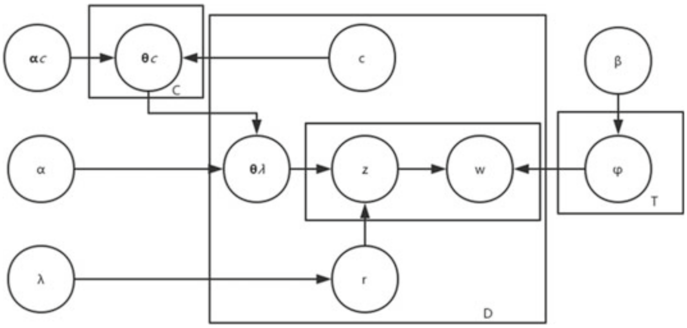


Fig. 3 Bayesian network diagram of MB-LDA model

Among them, r and c are used to represent the forwarding relationship and the contact relationship, and the probability distribution of θ in the entire microblog text set is as follows:

$$\mathrm{p}(\theta|\alpha, \alpha_c, \mathrm{c}) = \mathrm{p}(\theta_c|\alpha_c)_c^x \mathrm{p}(\theta_d|\alpha)^{1-x_c} \tag{2}$$

When we want to start mining microblog topics, the MB-LDA topic model extracts the relationship between a certain topic and a word from the Dirichlet distribution of β, and records it as φ. Whenever the MB-LDA topic model mines a microblog, you need to use @ to judge the relationship between contacts and contacts. Since @ appears in the middle of the microblog blog post, the MB-LDA topic model cannot be used to determine whether @ here wants to establish a relationship with the contact or other meaningless relationship, so only consider the case where @ appears at the beginning of the blog post. At this time, the relationship between the contact c and each topic is sampled from the Dirichlet distribution of α_c and recorded as θ_d. The joint probability distribution of all the words in a certain microblog and the topics it belongs to is shown in Eq. (3):

$$\begin{aligned}\mathrm{p}(\mathrm{w}, \mathrm{z}|\lambda, \theta, \beta) &= \mathrm{p}(\mathrm{r}|\lambda)\mathrm{p}(\mathrm{z}|\theta)\mathrm{p}(\mathrm{w}|\mathrm{z}, \mathrm{p}) \\ &= \mathrm{p}(\mathrm{r}|\lambda)\mathrm{p}(\mathrm{z}|\theta_d)^{1-r}\mathrm{p}\big(z|\theta_{d_{RT}}\big)^R\mathrm{p}(\mathrm{w}|\mathrm{z}, \beta)\end{aligned} \tag{3}$$

In summary, the generation process of the MB-LDA topic model is shown in Algorithm 1:

Algorithm 1

```
For each topic k ∈ {1,2,3 ….$T$} $do$
  Select topic-word distribution $\varphi_k \sim Dir(\beta)$
End for
For each word $w_{dn}$ do
  Use RT to judge relationship
   If r = 1
      Select topic-word distribution $z_{dn} \sim Multi(\theta_{d_{RT}})$
      Otherwise $z_{dn} \sim Multi(\theta_d)$
  Otherwise, select topic-word distribution
  $z_{dn} \sim Multi(\theta_d)$
  Select topic-word distribution $w_{dn} \sim Multi(\varphi_{d_{RT}})$
 End for
End for
```

4.2 Derivation of MB-LDA Topic Model

The MB-LDA topic model is derived using Gibbs Sampling, which is a fast and efficient MCMC (Markov chain Monte Carlo) sampling method that is used to derive complex probability distributions by iterative sampling and is mostly used for solving Bayesian diagram models. The process of deriving the MB-LDA topic model is as follows:

The joint probability distribution of all words in a text and the subject to which they belong is shown in formula (4).

$$P(w, z|\alpha, \beta) = P(w|z, \beta)P(z|\alpha, \beta) \int P(z|\theta)P(\theta|\alpha)\mathrm{d}\theta \int P(w|z, \varphi)P(\varphi|\beta)\mathrm{d}\varphi \tag{4}$$

Exploitation of formula (4) using the Euler formula

$$\begin{aligned} P(w|z, \beta) &= \left(\frac{\Gamma(V\beta)}{\Pi_v \Gamma(\beta)}\right)^T \prod_{j=1}^{T} \frac{\Pi_v \Gamma\left(n_{j,v} + \beta\right)}{\Gamma\left(n_{j,1} + V\beta\right)} \\ P(z|\alpha) &= \left(\frac{\Gamma(T_\alpha)}{\Pi_j \Gamma(\alpha)}\right)^T \prod_{d=1}^{D} \frac{\prod_j \Gamma\left(n_{d,j} + \alpha\right)}{\Gamma\left(n_{d,\cdot} + T_\alpha\right)} \end{aligned} \tag{5}$$

The following a posteriori distributions were then sampled using Gibs Sampling.

$$\mathrm{p}(z_i = \mathrm{j}|\mathrm{w}, z_{-i}, \alpha, \beta) = \frac{\mathrm{p}(\mathrm{z}, \mathrm{w}|\alpha, \beta)}{\mathrm{p}(z_{-i}, \mathrm{w}|\alpha, \beta)} \propto \frac{n_{jv} + \beta - 1}{n_{i+}V_\beta - 1} * \frac{n_{dj} + \alpha - 1}{n_d + T_\alpha - 1} \tag{6}$$

and sampling of all subjects resulted in stable sampling results. The final result obtained for the hidden variable is

$$\theta_d = \frac{n_{dj} + \alpha - 1}{n_d + T_\alpha - 1} \tag{7}$$

$$\varphi_z = \frac{n_{jv} + \beta - 1}{n_j + V_\beta - 1} \tag{8}$$

The complete training process for the MB-LDA topic model is shown in Algorithm 2:

```
Algorithm 2
1. Integrate all microblog blog posts and get user texts.
2. Initialize a topic z ∈ $\{1,2,3 \ldots. T\}$ for each word w ∈ $\{1,2,3 \ldots. V\}$ in all texts.
3. Update the topic-word distribution of all texts, that is, sample each word.
4. Iterate repeatedly until Gibbs sampling converges.
5. According to the convergence result, the topic-word probability distribution is obtained.
6. According to the obtained topic-word probability distribution, each topic word is sorted from high frequency to low frequency.
```

Therefore, in summary, compared with the Twitter-LDA model, the MB-LDA model has made two improvements: First, introduce a section on user relationships and textual relationships, and make the topic of the blog post explicit as a context for blog posts where the topic is not clear. Second, the introduction of additional topics for invalid microblogging, the existence of invalid microblogging affects the microblogging blog topic mining, the introduction of additional topics can filter the impact of invalid microblogging. Moreover, the MB-LDA model can set different additional topics for each user to uncover the invalid tweets against the user.

4.3 Implementation of the MB-LDA Topic Model

This paper uses Python to build and experiment on the MB-LDA topic model.

1. Use GooSeeker to crawl all valid microblog text, integrate it into an Excel document, and clean up and preprocess the data.
2. Prepare the Document-Term matrix, that is, convert the corpus into a Document-Term matrix. The corpus consists of all files. The working principle of the MB-LDA topic model is to find repeated word patterns in the entire Document-Term matrix. Python provides many good libraries for text mining tasks. Among them, “genism” is a better library for processing text data.

3. Formally build the MB-LDA topic model. Its steps are mainly to create an MB-LDA object and use the Document-Term matrix for training. The training requires some of the above hyperparameters. The Genism module allows the MB-LDA topic model to be estimated based on the training corpus, and inferences about topic distribution from new documents.
4. Output results. In the modeling of the MB-LDA topic model, it is assumed that text clustering requires 8 topics, and each topic has 10 keywords.

5 Experimental Design

5.1 Data Collection

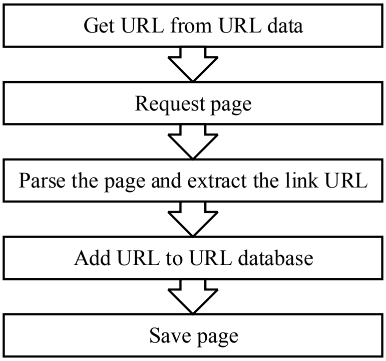
There are two ways to obtain Sina microblog data, one is to use Sina microblog open platform API, and the other is to use web crawlers. The Sina microblog open platform API is officially provided by Sina microblog, but there are many restrictions on its use. To obtain all the comments of a blog post, you must obtain the authorization of the user of these comments. Obviously this is not feasible, so this paper uses Web crawlers to get data. The principle of the web crawler is to obtain webpage data according to the web URL in the format of HTML and then parse the obtained HTML data to obtain the required information from it.

This paper uses the web crawler tool Goo Seeker, which provides a microblog collection toolbox for Sina microblog, which can easily collect "detailed information of microblog bloggers", "microblog blogger homepage content", and "microblog repost/comment information", "fan and follower information" and other data.

The principle of the web crawler is shown in Fig. 4.

Through the web crawler Goo Seeker, this paper acquired thousands of the latest microblog data as well as the forwarding and comments of these microblogs. The data were organized according to users and then filtered. The number of posts or comments was not enough to filter out. Too few posts or comments were not enough

Fig. 4 How web crawlers work



to support the experiment in this paper. Preliminary processing of the remaining data, removing multimedia data, emoji data, hyperlink data, etc. from the blog post. In the end, 3311 valid data were obtained. Some of these data are shown in Table 4.

5.2 Experimental Results and Analysis

5.2.1 Experimental Results Display

In this paper, when the MB-LDA topic model is used to excavate the topic of emergency information related to "Typhoon Mangkhut", 8 topics are clustered, and each topic has 10 keywords. The results of the excavated topics are shown in Table 5.

Based on the output of the MB-LDA topic model, we can analyze the first eight topics related to emergency rescue, blessing, Yangchun disaster, a delivery man in distress, flight delay, the everlasting stone blown down in Dameisha, emergency measures and weather warning. Also, by clustering Topic #1, Topic #2, Topic #3, and Topic #4, keywords with top 30 words frequencies under that topic are tapped. As shown in Figs. 5, 6, 7 and 8, the left side of each image is the four topics in the cluster and the right side is the top 30 keywords under that topic.

5.2.2 Performance Analysis Based on Perplexity

Perplexity is a common indicator for evaluating the performance of statistical language models. In this experiment, the index of perplexity is used to quantitatively evaluate the MB-LDA topic model and the traditional LDA topic model, and compare and analyze their performance. The lower the calculated Perplexity value, the lower the blur degree of the model and the better the performance. The calculation formula of Perplexity is shown in (9):

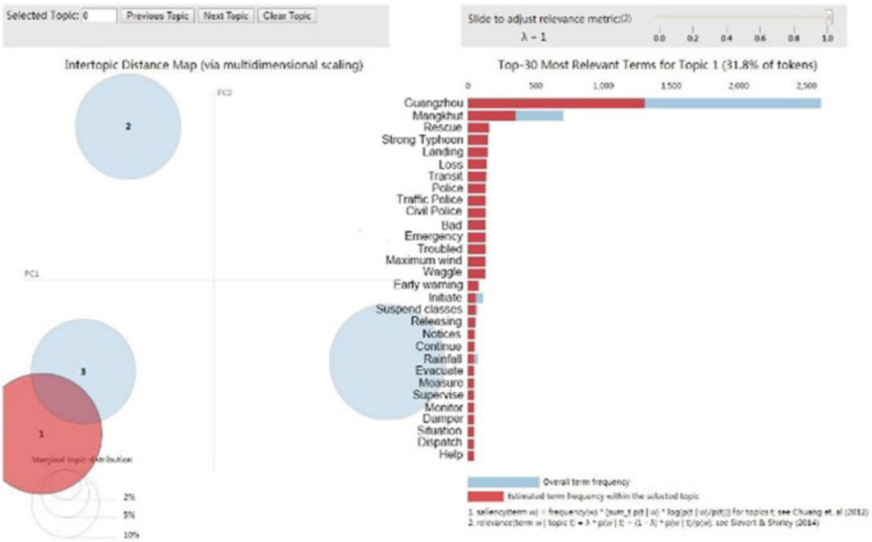
$$Perplexity(\mathrm{W}) = \exp\left\{-\frac{\sum_{m=1}^{M}\log p(\vec{w}_m)}{\sum_{m=1}^{M} N_m}\right\} \tag{9}$$

Table 4 Some data display of web crawlers

Keyword	Some blog posts show	Number of reposts
Typhoon Mangkhut	After the wind king “Mangkhut” retired, whether the fierce “Mangkhut” should be delisted was hotly discussed	0
Typhoon Mangkhut	Mangkhut crossing, Hong Kong Disneyland reopened within 1 day after cleaning. Netizen: It was repaired with magic	163
Typhoon Mangkhut	Recently, the restoration work after the typhoon “Mangkhut” raged is underway, and more than 100 Su-Ning stores in Guangzhou, Shenzhen and other regions have also responded quickly	16
Typhoon	In the evening of the third day after the landing of the “Mangkhut”, the rescue personnel of the Guangdong Power Grid Huizhou Daya Bay Bureau was still on the scene to replace the seriously damaged equipment to prepare for the restoration of power supply	7
Typhoon	Guangzhou traffic police responded to “Typhoon early in the morning to copy cards”: the record content failed and apologized	1
Typhoon	Affected by typhoon Mangkhut, some trains were still suspended in the southeast coastal areas of Shenzhen and Guangzhou on September 18 and 19	0
Mangkhut	Because the name of a typhoon is recycled, Mangkhut, if removed from the list, refers specifically to the massive hurricane that made landfall in Taishan, China, this year	29
Mangkhut	“Mangkhut” affects less blood storage, Shenzhen Blood Center urges loving people to actively donate blood	3
Mangkhut	Typhoon Mangkhut made landfall off the coast of Taishan in Guangdong province with a maximum wind force of 14 near the center	13
Typhoon Mangkhut	Typhoon “Mangkhut” disaster: 5 people died, 1 person is missing, direct economic loss of 5.2 billion	56
Typhoon Mangkhut	Some tree experts pointed out that Hong Kong and Shenzhen will enter the dangerous period of tree collapse in the coming week. Due to experience, many accidents that caused serious casualties occurred after the typhoon	20
Typhoon Mangkhut	After the typhoon “Mangkhut” struck, the Pearl River water poured back, and many stalls of the warehouses of the Haidui Dry Goods Wholesale Market on Guangzhou Yide Road were flooded. A large number of soaked mushrooms and cloud ears were discarded in nearby streets, and they smelled bad	4

Table 5 MB-LDA topic model output results

Topic	Keywords
Topic #1	Guangzhou, Mangkhut, Rescue, Strong Typhoon, Landing, Loss, Government, Police, Traffic Police, Civil Police
Topic #2	Mangkhut, Typhoon, Bodhisattva, Avalokitesvara, Bless, Rise, Protect, Night, Peace, Cheer
Topic #3	Mangkhut, Waterlogging, Typhoon, Warning, Yangchun, Part, Yangjiang, Residents, Exceeded, Water Leve
Topic #4	Deliveryman, Typhoon, Mangkhut, Branches, Express, Brother, Teenager, All-Around, Break, Wind and Rain
Topic #5	Typhoon, Mangkhut, Flight, Yunnan, Man, Influence, Hong Kong, Netizen, Guangdong, Delayed
Topic #6	Mangkhut, Typhoon, Enduring, Oath, Netizens, Blowdown, Shenzhen, Dameisha, Stone, Everlasting
Topic #7	Typhoon, Mangkhut, Weather, Observatory, Work, Guangdong, Serious, Landing, Strong Typhoon, Forecast
Topic #8	Typhoon, Mangkhut, Close, Doors And Windows, Shenzhen, Damper, Landing, Food, Know, Swing


Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 5 Topic # 1 top 30 keywords

Among them, W represents the corpus of M texts, N_m represents the number of words contained in the text m, p($\vec{w}_m$) represents the probability of generating the text m, p($\vec{w}_m$)'s calculation formula is shown in (10):

$$\mathrm{p}(\vec{w}_m) = \Pi_{w_i \in m} \Sigma_{z \in T}\, p(W_i|z)p(z|m) \tag{10}$$

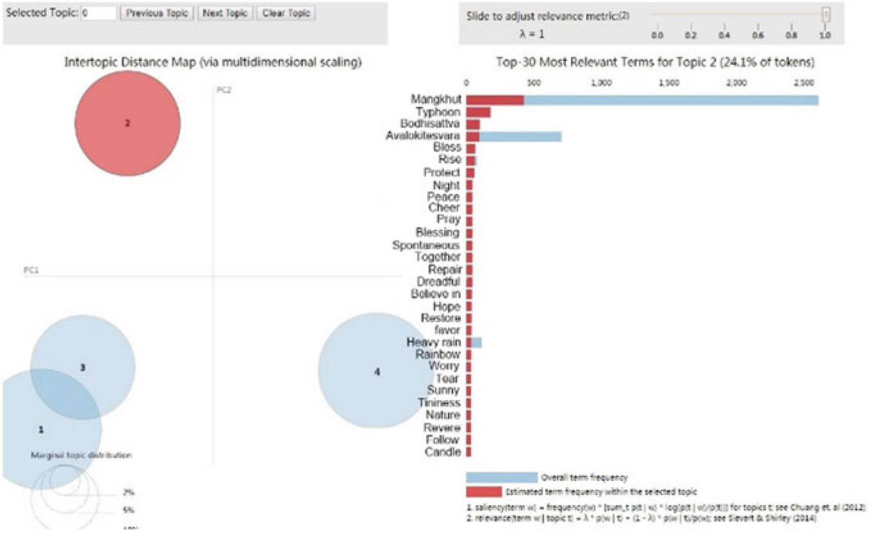


Fig. 6 Topic # 2 top 30 keywords

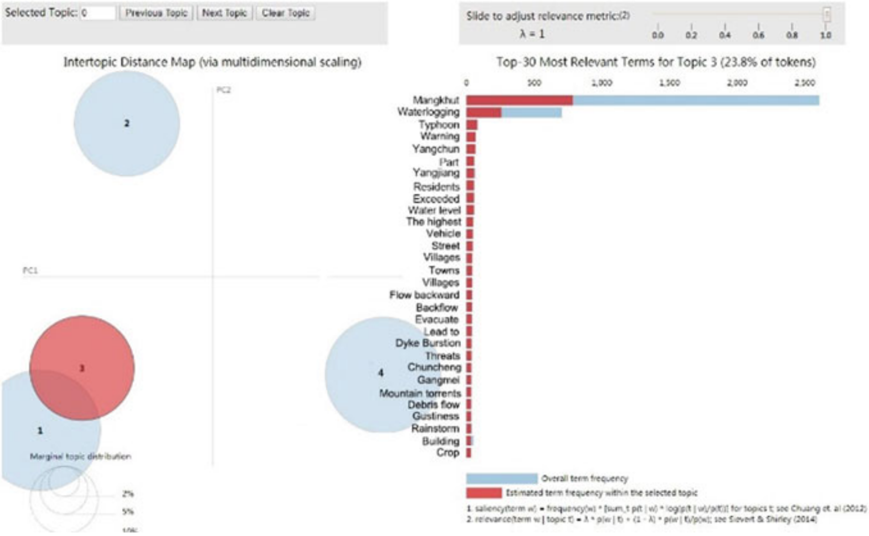
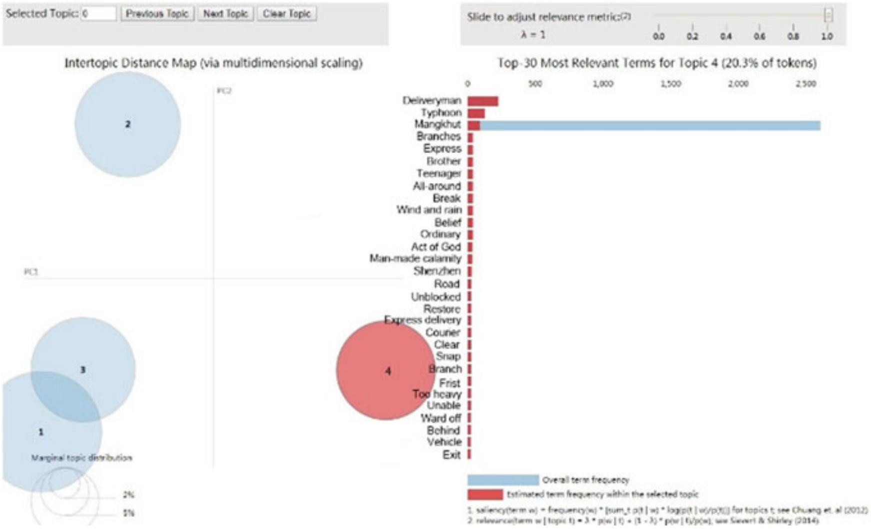


Fig. 7 Topic # 3 top 30 keywords

The setting of the number of topics r will directly affect the performance of the training model. If the number of topics is too large or too small, it will produce a bad analysis effect. Different corpora correspond to different optimal topics. It is necessary to set different topics to train the corpus, find the corresponding confusion, and select the topic number r with the lowest Perplexity as the optimal parameter



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

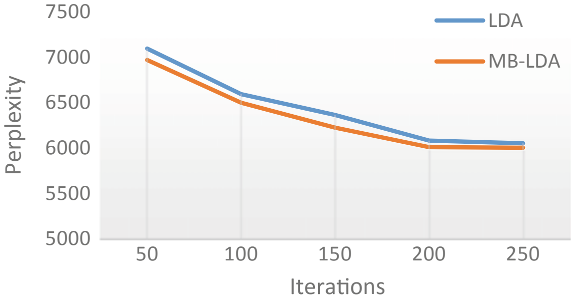
Fig. 8 Topic # 4 top 30 keywords

value. In the experiment, the number of topics r = 5, 10, 15, 20, 25, 30, 35, 40 was selected. The LDA topic model and MB-LDA topic model were used to train the corpus 50 times.

The perplexity does not change monotonously with the number of topics. When the number of topics is 20, the MB-LDA topic model and the LDA topic model have the lowest perplexity, that is, the best performance. Therefore, the optimal number of topics corresponding to the experimental microblog corpus is 20. Under the same corpus and parameter settings described above, namely α = 20/T, β = 0.01, T = 20, the corpus is trained according to different iteration times. The performance comparison between the MB-LDA topic model and the LDA topic model is obtained as shown in Fig. 9.

According to the comparison of experimental results, under the same environment, as the number of training iterations continues to increase, the Perplexity of the two

Fig. 9 Comparison of the perplexity of two models



models becomes lower and lower. When the number of iterations is more than 150, the Perplexity tends to be stable, and the model reaches the convergence state. More importantly, the Perplexity of the MB-LDA topic model is always lower than the traditional LDA topic model, which shows that the improved MB-LDA model is superior to the standard LDA model in performance, and is more suitable for mining the emergency information topic of Chinese microblog.

5.2.3 Experimental Effect Analysis

In the experiment, the MB-LDA topic model and the traditional LDA model were used to perform topic mining on the microblog dataset, and the effects of the topic information mined by these two models were compared and analyzed. The number of sampling iterations was set to 50. According to the probability distribution of the words under the 8 topics output by the model, the top 10 words with the highest distribution probability are selected as the keywords of the topic and their effects are displayed. At the same time, the topic ranking algorithm is used to rank the importance of topics. From the analysis and comparison of the experimental results, it can be seen that the MB-LDA topic model is superior to the traditional LDA topic model for Chinese microblog topic mining, which can more effectively mine topic information and enhance the readability and independence of the topic.

The MB-LDA topic model based on the user dimension can train the user-topic probability distribution. The higher the probability of topic distribution among users, the more interested the user is about the topic. Therefore, by sorting the topics in descending order according to the user-topic distribution, the topics of interest to each user can be obtained.

6 Conclusion

With the rapid development of Internet technology and the increasing popularity of social networks, Microblog, as a form of social network that has emerged in recent years, has the characteristics of convenience, interaction and real-time, attracting many users, and has become an important information-sharing platform. Therefore, it is of great significance for the government to grasp the public opinion to develop effective methods to quickly excavate valuable emergency information topics from massive, fuzzy and obscure microblog text data.

The traditional LDA topic model is suitable for long text topic mining, but not for short text data sets such as microblog. Because the content of microblog is limited to 140 words and contains less semantic information, and the content of the microblog content published by each user involves a relatively concentrated subject area, this paper introduces the user dimension to integrate the microblog data of the same user to form a long text with rich semantic information establishes a four-layer LDA model of “text-user-topic-word”. At the same time, given the proliferation of microblog marketing information, the randomization of microblog content, and the problem of a lot of noise, this paper uses Chinese word segmentation technology to pre-treat the crawled emergency information microblog data set to eliminate interference information to improve the accuracy of microblog topic mining. In the end, an MB-LDA topic model suitable for Chinese emergencies for emergency information topic mining is proposed. Experiments have proved that the MB-LDA topic model is superior to the traditional LDA topic model in performance and is more suitable for Chinese emergency information topic mining on the microblog.

Acknowledgements This work was partially supported by Beijing Logistics Informatics Research Base and Intelligent Emergency Logistics Linkage System of Public Emergencies in Beijing under Grant 18JDGLB019.

References

1. W.X. Zhao, J. Jiang, J. Weng et al., Comparing twitter and traditional media using topic models, in *Proceedings of Advances in Information Retrieval—33rd European Conference on IR Research (ECIR)* (2011), p. 338
2. H. Tu, J. Ding, An efficient clustering algorithm for microblogging hot topic detection, in *Proceedings of International Conference on Computer Science & Service System (CSSS)* (2012), pp. 738–741
3. F. Zhao, Y.J. Zhu, H. Jin, L.T. Yang, A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Future Gener. Comput. Syst. Int. J. Escience **65**, 196–206 (2016)
4. Q. Wen, M.S. Qiang, B.Q. Xia, N. An, Discovering regulatory concerns on bridge management: an author-topic model based approach. Transp. Policy **75**, 161–170 (2019)
5. J. Wang, L. Li, F. Tan, Y. Zhu, W. Feng, Detecting hotspot information using multi-attribute based topic model. PLoS ONE **10**(10), 1–16 (2015)
6. Z.A. Zhi, J.P. Zeng, Optimal selection method for LDA topics based on degree of overlap and completeness. Comput. Eng. Appl. **55**(12), 155–161 (2019)
7. J. Yu, L. Qiu, ULW-DMM: an effective topic modeling method for microblog short text. IEEE Access **7**, 884–893 (2019)
8. T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. **25**(24), 919–931 (2013)
9. C. Andrew, C. Arie, S. Anthony, R. Jacek, Earthquake: twitter as a distributed sensor system. Trans. GIS **17**(1), 124–147 (2013)
10. X.H. Han, J.L. Wang, Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in China. Isprs Int. J. Geo-Inf. (2019). https://doi.org/10.3390/ijgi8040185
11. T.H. Ma, J. Li, X.N. Liang, A time-series based aggregation scheme for topic detection in microblog short texts (2019). https://doi.org/10.1016/j.physa
12. L.L. Fu, Y.H. Dong, Research on internet search data in China's social problems under the background of big data. J. Logistics Inf. Serv. Sci. **5**(2), 55–67 (2018)
13. J. Bian, Y. Yang, H. Zhang, T. Chua, Multimedia summarization for social events in microblog stream. IEEE Trans. Multimedia **17**(2), 216–228 (2015)
14. P. Wu, X. Li, H.S.D. Shen, Social media opinion summarization using emotion cognition and convolutional neural networks. Int. J. Inf. Manage. **51**, 1–15 (2020)
15. Y. Xi, J.Y. Zhao, W.J. Liu, The modeling and analyzing methods for enterprise micro-blog topics dissemination supernetwork based LDA. Chin. J. Manag. **15**(3), 434–441 (2018)

16. Y.R. Chen, W. Chen, English translation of long traditional Chinese medicine terms a corpus-based study. Terminology **24**(2), 181–209 (2018)
17. M. Memon, Y. Lu, P. Chen, A. Memon et al., An ensemble clustering approach for topic discovery using implicit text segmentation. J. Inf. Sci. (2020). https://doi.org/10.1177/0165551520911590
18. U. Claude, Predicting tourism demands by google trends: a hidden markov models based study. J. Syst. Manage. Sci. **10**(1), 106–120 (2020)
19. J.F. Ren, C.M. Ye, F. Yang, A novel solution to JSPS based on long short term memory and policy gradient algorithm. Int. J. Simul. Model. **19**(1), 156–168 (2020)
20. J. Zhang, D. Chang, Semi-supervised patient similarity clustering algorithm based on electronic medical records. IEEE Access **7**, 90705–90714 (2019)
21. Q. Guan, S. Deng, H. Wang, Chinese stopwords for text clustering: a comparative study. Data Anal. Knowl. Discov. **1**(03), 72–80 (2017)
22. J. Li, S.X. Pan, L. Huang, X. Zhu, A machine learning based method for customer behavior prediction. Tehnicki Vjesnik-Technical Gazette **26**(6), 1670–1676 (2019). https://doi.org/10.17559/TV-20190603165825
23. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik-Technical Gazette **25**(2), 528–535 (2018)
24. L.M. Wang, Z.Y. Hao, X.M. Han, R.H. Zhou, Gravity theory-based affinity propagation clustering algorithm and its applications. Tehnicki Vjesnik-Technical Gazette **25**(4), 1125–1135 (2018)
25. L.J. Hong, B. Davison, Empirical study of topic modeling in Twitter, in *Proceedings of the First Workshop on Social Media Analytics (SOMA'10)* (2010), pp. 80–88
26. D. Quercia, H. Askham, J. Crowcroft, TweetLDA: supervised topic classification and link prediction in Twitter, in *Proceedings of the 3rd Annual ACM Web Science Conference* (2012), pp. 247–250

Application of Object—Attribute Space Segmentation in Bidding Activities

Yijie Yin, Yuwen Huo, and Yaoyu Hu

Abstract In this paper, the object-attribute space segmentation method is used to segment the high-dimensional sparse matrix of regulation-bidding process in bidding activities, and the correlation model of regulation-bidding process is constructed to improve the compliance inspection efficiency in bidding activities. This paper analyzes the main steps and key problems of the algorithm, and finally applies the algorithm to carry out case analysis. The results show that the algorithm can effectively reduce the size and dimension of data and improve the efficiency of data mining.

Keywords Bidding · Object-attribute space segmentation · Data mining · Clustering analysis

1 The Introduction

The 4 million tons/year coal indirect liquefaction project of Shenning group is the first large-scale commercial project in the indirect liquefaction route of coal in China's coal-to-liquids industry and the largest coal-to-liquids project in the world. Bidding in projects is a comprehensive form of economic responsibility system to promote competition in the field of capital construction. There are a large number of business operations and documents in the bidding activities, and all the operations must not violate the laws promulgated by the state, so compliance judgment becomes an important work content in the bidding activities [1, 2].

Y. Yin (✉) · Y. Huo · Y. Hu
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: daisy5267@foxmail.com

Y. Huo
e-mail: 13698896904@163.com

Y. Hu
e-mail: 13707537@chnenergy.com.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_45

The regulation-bidding correlation model is to realize the structured description and storage of the regulation-bidding correlation by modeling the laws and the bidding process unit, assist the related person to complete the compliance determination of the bidding activities, and control the bidding risk [3]. The research object of the regulation-bidding association model is the relevant laws and regulations of bidding and the bidding process. The purpose of establishing the regulation-bidding association model is:

1. automatically identify the scope of the law for a given law, that is, all bidding process/document units subject to the law;
2. for a given unit of bidding process/document, the legal unit that constrains the unit is automatically built.

There are several steps to establish the regulation-bidding association model [4]:

1. divide the bidding process unit and construct the structural model of the bidding process unit;
2. divide the units of bidding documents and build a structural model of the units of bidding documents;
3. construct the structural model of laws and regulations and the structural model of laws and regulations;
4. analyze the constraint relationship between the law and the bidding process unit during the execution of bidding activities, and construct the correlation model of regulation-bidding activities.

The specific process is shown in Fig. 1.

When constructing the correlation model of regulation-bidding activities, the bidding process unit is taken as the object and the regulations as the attribute to form the object-attribute matrix. Due to the fact that in the bidding activities, there are generally few bidding process units that apply the same law, but many laws are needed for a bidding process unit, so the correlation matrix of regulation-bidding process unit is a high-dimensional sparse matrix [5]. The spatial partition of the

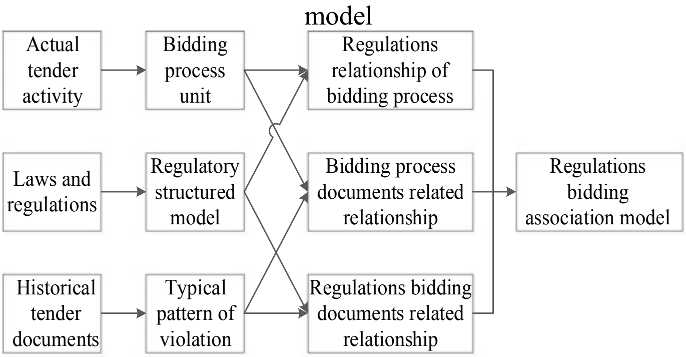

model
Actual tender activity
Bidding process unit
Regulations relationship of bidding process
Laws and regulations
Regulatory structured model
Bidding process documents related relationship
Regulations bidding association model
Historical tender documents
Typical pattern of violation
Regulations bidding documents related relationship

Fig. 1 Construction process of regulation bidding association model

correlation matrix of the regulation-bidding process unit is essentially the spatial partition of the sparse object-attribute matrix with high dimensions.

2 Related Work

The object—attribute space segmentation method is used to segment the high-dimensional sparse regulation-bidding process element correlation matrix [6], and the large scale data is simplified from two aspects of data size and attribute dimension before the formal data mining, so that the classical data mining algorithm can get better mining results. The following two concepts (object cohesion and object aggregate cohesion) are introduced.

2.1 *Object Cohesion*

Object cohesion: suppose there are n objects, denoted as $O = \{O_1, O_2, \ldots, O_n\}$, there are m attributes to describe each object, denoted as $A = \{A_1, A_2, \ldots, A_m\}$. The non-zero attribute set of Object O_i (i = 1, 2, …, n) (the attribute set of object attribute characteristic value takes 1) is OA_i (i = 1, 2, …, n). The number of attributes is denoted as $|OA_i|$ (i = 1, 2, …, n), and the cohesion of two objects O_i and O_j is defined as:

$$OC\left(O_i, O_j\right) = \frac{\left|OA_i \cap OA_j\right|}{\left|OA_i \cup OA_j\right|} \quad (\mathrm{i,\ j} = 1, 2, \ldots,\ \mathrm{n\ and\ i} \neq \mathrm{j})$$

If and only if i = j, $OC(O_i, O_j) = 1$, the cohesion of the object and itself is 1. Object cohesion reflects the degree of similarity between two objects.

2.2 *Object Aggregate Cohesion*

Object aggregate cohesion: suppose set *X* and set *Y* are the disjoint object subsets in *O*, the cohesion degree of object set *X* and set *Y* is defined as the maximum object cohesion degree of objects in set *X* and set *Y*, that is:

$$\mathrm{OAC(X, Y)} = \max_{\mathrm{O_i \in X, O_j \in Y}}\{\mathrm{OC(O_i, O_j)}\} \quad (\text{i indicates j})$$

The object aggregate cohesion measures the cohesiveness of all objects in a collection due to their attributions. The higher the object aggregate cohesion is, the more similar the objects in or between the sets are, the stronger the cohesion is, and the more likely they are to be clustered into a class. Conversely, the less similar the objects are, the less likely they are to be clustered into a class.

3 Principle and Procedure of Algorithm

3.1 Algorithm Principle

The object-attribute space segmentation method is to pre-cluster the high-dimensional sparse data according to the characteristics of object-attribute in the process of data mining, and divide the original data into several sub-intervals. After partition, the size and dimension of the data will have a certain degree of reduction. Finally, we only need to connect each small subspace to form the final clustering result [7].

The basic idea of object-attribute space segmentation: In the first stage, based on the idea of agglomerating hierarchical clustering algorithm, the similarity between object sets is used as the measurement index to cluster the objects with non-zero attribute approximation, and the whole object-attribute space is divided into several object-attribute subsystems [8]. In the second stage, the subsystem obtained in the first stage is pruned to reduce the properties with little influence in each subsystem [9].

3.2 Algorithm Steps

- Input: object—attribute relation matrix, the threshold value of cohesion between object sets α, the threshold value of attribute reduction β.
- Output: object set (subsystem) and its corresponding set of attribute.
- Steps:

Step 1. In the initial data, each object creates a collection separately, denoted as $X_i(0)$, $i \in \{1, 2, \ldots, n\}$

Step 2. Calculate the similarity of set $X_1(0)$ and $X_2(0)$ and define it as *OAC* $(X_1(0), X_2(0))$, if *OAC* $(X_1(0), X_2(0)) \geq \alpha$, then merge $X_1(0)$ and $X_2(0)$. The merged class is denoting as $X_1(1)$. If *OAC* $(X_1(0), X_2(0)) < \alpha$, respectively denoting $X_1(0)$ and $X_2(0)$ as class $X_1(1)$ and class $X_2(1)$. Call the number of classes c.

Step 3. Loop calculation $X_3(0)$ and $X_i(1)$ $(i \in \{1, 2, \ldots, c\})$. If *OAC* $(X_3(0), X_i(1)) \geq \alpha$, then merge class $X_3(0)$ and class $X_i(1)$, and exit the loop. The merged class is still referred to as $X_i(1)$ and the number of classes c does not change. If $i \in \{1, 2, \ldots, c\}$, *OAC* $(X_3(0), X_i(1))$ are all less than α, then $X_3(0)$ is taken as a new class which denoted as $X_{c+1}(1)$. The number of classes c = c + 1.

Step 4. Operate Xi(0) i ∈ {4, 5, …, n} in turn in step 3.

Step 5. Consider $X_c(1)$ as a new object to be clustered, and proceed with steps 2 through steps 5. Until the resulting class is no longer changing.

Step 6. Attribute reduction is carried out within the class for the formed classes. Calculate the number of times that all objects value attributes in each class. If the ratio of the number of times an attribute takes a value to the total number of objects in that class is less than β, the attribute is considered to have less influence in the future data mining and the attribute is removed from the class.

Step 7. Output the completed set of objects and their corresponding set of attribution.

For the convenience of understanding, a flow chart shown as Fig. 2 is used to describe the calculation process of the algorithm.

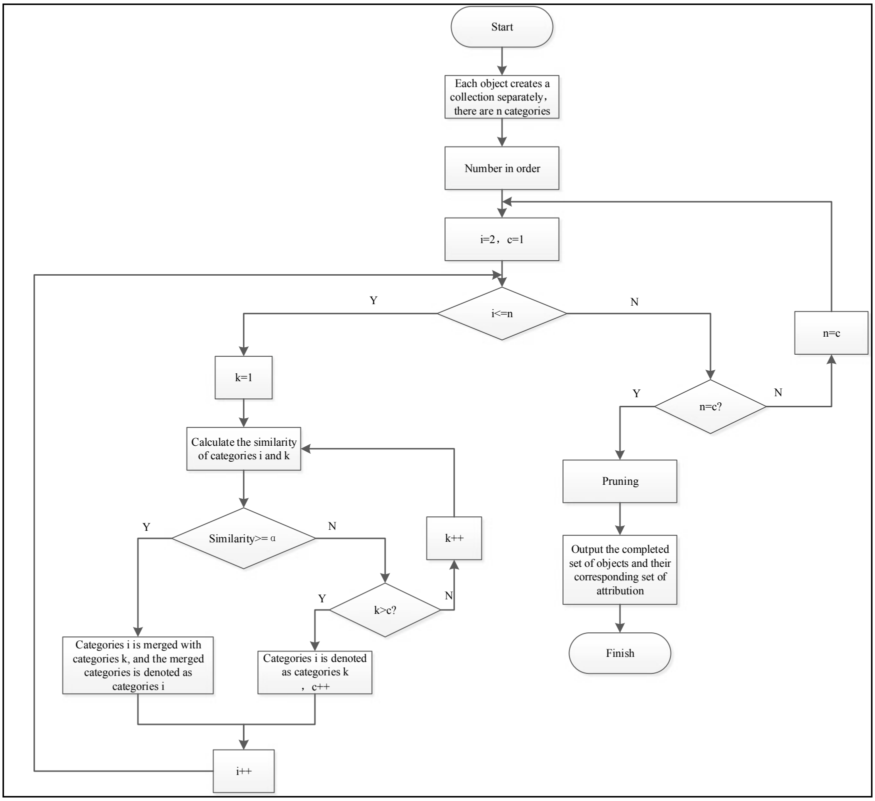

Start
Each object creates a collection separately, there are n categories
Number in order
i=2，c=1
i<=n
Y
N
k=1
Calculate the similarity of categories i and k
Similarity>= α
Y
N
k++
k>c?
Y
N
Categories i is merged with categories k, and the merged categories is denoted as categories i
Categories i is denoted as categories k ，c++
i++
n=c?
Y
N
n=c
Pruning
Output the completed set of objects and their corresponding set of attribution
Finish

Fig. 2 Algorithm flow chart

4 The Example Analysis

4.1 Problem Background and Data

Shenning group 4 million tons/year indirect coal liquefaction project laws—bidding project association model is mainly divided into laws module, bidding process module and association module. According to the existing documents with legal effect and the regulations issued by the state on the gasification unit of coal-to-oil project with a capacity of 4 million tons/year, a regulation module is constructed to form 14 law units as shown in Table 1. According to the process of bidding activities with sequential relationship in the bidding process, the bidding process module is constructed to form 51 bidding process units as shown in Table 2.

The high-dimensional sparse matrix is divided into space with the bidding process unit taken as the object and the rule as the attribute. Table 3 shows the data of regulation-bidding process with 51 objects and 14 attributes.

Table 1 Laws and regulations related to bidding activities

Name of laws and regulations	Regulation number
Law of the People's Republic of China on tendering and bidding	1
Regulations for the implementation of the bidding and tendering law of the People's Republic of China	2
Law of the People's Republic of China on government procurement	3
Regulations on the implementation of the government procurement law of the People's Republic of China	4
Regulations on bidding and procurement management of shenhua group company (revised in 2016)	5
Methods for bidding and tendering of goods for construction projects	6
Methods for survey and design bidding and bidding of engineering construction projects	7
Methods of bidding and tendering for construction projects	8
Measures for handling complaints about bidding and tendering activities of construction projects	9
Interim provisions on bid assessment committees and bid assessment methods	10
Interim measures for the administration of bid assessment experts and bid assessment expert database	11
Provisional measures for the promulgation of the tender announcement	12
Electronic tendering and bidding method	13
Implementation rules for shenhua group's material purchase management (revised in 2017)	14

Table 2 Division of bidding process units

Bidding method	Unit 1 of bidding process	Unit 2 of bidding process	Unit number of bidding process
Public bidding			
	Tender preparation stage		1
		Establish the delegate	2
		To compile the bidding plan	
		Preparation of bidding plan	3
	Prequalification stage		4
		Prepare prequalification documents	5
		Confirmation of prequalification documents	6
		Issue a prequalification notice	7
		Pre-qualification documents for sale	8
		Clarification of prequalification documents	9
		To receive eligibility documents	10
		Set up the appraisal committee	11
		prequalification	12
		Prequalification report	13
		Confirm the list of pre-qualified bidders	14
	The bidding stage		15
		Preparation of bidding documents	16
		Confirmation of bidding documents	17

(continued)

Table 2 (continued)

Bidding method	Unit 1 of bidding process	Unit 2 of bidding process	Unit number of bidding process
		Issue of tender notice (post-qualification examination)/issue of invitation to bid to bidders who have passed the pre-qualification examination (pre-qualification examination)	18
		Tender documents for sale	19
		Survey/preparatory meeting/q&a	20
		Selection of experts from the bid assessment committee	21
		Acceptance of qualified tender documents	22
		The bid opening	23
		The bid assessment	24
		The evaluation to clarify	25
		The bid evaluation report	26
		The tenderer shall confirm the bid assessment results	27
		Publicity of winning candidates	28
		Complaint and objection handling	29
		Issue the notification of winning/losing the bid	30
		Deposit refund/bid winning service fee	31
Invitation to tender			
			32
		Establish the delegate	
		To compile the bidding plan	33

(continued)

Table 2 (continued)

Bidding method	Unit 1 of bidding process	Unit 2 of bidding process	Unit number of bidding process
		Preparation of bidding plan	34
	The bidding stage		35
		Preparation of bidding documents	36
		Confirmation of bidding documents	37
		Send out invitation for bid to potential bidders	38
		Tender documents for sale	39
		Survey/preparatory meeting/q&a	40
		Selection of experts from the bid assessment committee	41
		Acceptance of qualified tender documents	42
		The bid opening	43
		The bid assessment	44
		The evaluation to clarify	45
		The bid evaluation report	46
		The tenderer shall confirm the bid assessment results	47
		Publicity of winning candidates	48
		Complaint and objection handling	49
		Issue the notification of winning/losing the bid	50
		Deposit refund/bid winning service fee	51

4.2 Algorithm Practice Process

Step 1. Set up a set for each object denoted as $X_{\mathrm{i}}(0)$, i ∈ {1, 2, …, 51};

Table 3 Values of 51 objects and 14 attributes

Nummber of bidding process	Number of used laws
1	2, 3, 4, 5, 6, 7, 14
2	1, 2, 3, 4, 5, 6, 7, 8, 14
3	2, 3, 4, 5, 6, 7, 8
4	2, 6
5	2, 6, 8
6	6
7	2, 3, 4, 5, 6, 7, 8
8	7, 8
9	5, 7, 8
10	6, 8
11	3
12	2, 6, 8
13	2
14	6, 7, 8
15	1, 2, 4, 5, 6, 7
16	1, 2, 4, 5, 6, 7, 8, 14
17	5, 14
18	2, 3, 4, 5, 6, 7, 8, 14
19	2, 4, 5, 6, 7, 8, 14
20	2, 3, 4, 5, 6, 7, 8
21	2, 4, 5, 7, 8, 14
22	2, 5, 6, 7, 8, 14
23	1, 2, 3, 4, 5, 6, 7, 8, 14
24	1, 2, 3, 4, 5, 6, 7, 8, 14
25	1, 2, 6, 8
26	1, 2, 5, 6, 7, 8
27	1, 2, 3, 4, 5, 6, 7, 8
28	2, 3, 4, 5, 6, 7, 8, 14
29	1, 2, 3, 4, 5, 6, 8
30	1, 2, 3, 4, 6, 8
31	1, 2, 4, 6, 7, 8, 14
32	2, 3, 4, 5, 6, 7, 14
33	1, 2, 3, 4, 5, 6, 7, 8, 14
34	2, 3, 4, 5, 6, 7, 8
35	1, 2, 4, 5, 6, 7
36	1, 2, 4, 5, 6, 7, 8, 14
37	5, 14

(continued)

Table 3 (continued)

Nummber of bidding process	Number of used laws
38	2, 3, 4, 5, 6, 7, 8, 14
39	2, 4, 5, 6, 7, 8, 14
40	2, 3, 4, 5, 6, 7, 8
41	2, 4, 5, 7, 8, 14
42	2, 5, 6, 7, 8, 14
43	1, 2, 3, 4, 5, 6, 7, 8, 14
44	1, 2, 3, 4, 5, 6, 7, 8, 14
45	1, 2, 6, 8
46	1, 2, 5, 6, 7, 8
47	1, 2, 3, 4, 5, 6, 7, 8
48	2, 3, 4, 5, 6, 7, 8, 14
49	1, 2, 3, 4, 5, 6, 8
50	1, 2, 3, 4, 6, 8
51	1, 2, 4, 6, 7, 8, 14

Step 2. Calculate the cohesion of $X_1(0)$ and $X_2(0)$. *OAC* $(X_1(0), X_2(0)) = 7/9 > 65\%$. The combination condition is satisfied. So $X_1(0)$ and $X_2(0)$ are denoted as $X_1(1)$, and c = 1.

Step 3. The cohesion of $X_3(0)$ and $X_1(1)$ is *OAC* $(X_1(1), X_3(0))$ = Max (*OC* $(X_1(0)$, $X_3(0))$, *OAC* $(X_2(0), X_3(0)))$ = Max (6/8, 7/9) > 65%. So $X_3(0)$ is merged into $X_1(1)$, c = 1;

Step 4. The same as the method in step 3. *OAC* $(X_4(0), X_1(1)) < 65\%$ and they cannot be combined. $X_4(0)$ is denoted as $X_2(1)$. At this time the class increases, c = 2

Step 5. The cohesion of $X_5(0)$ and $X_1(1)$ is *OAC* $(X_5(0), X_1(1)) < 65\%$ and they cannot be merged. The cohesion of $X_5(0)$ and $X_2(1)$ is *OAC* $(X_5(0), X_2(1))$ = 2/3 > 65%. It is still denoted as $X_2(1)$ after merging. So $X_2(1) = \{X_4(0), X_5(0)\}$, c = 2

Similarly calculated, the first clustering results are 7 classes, respectively:

$X_1(1) = \{X_1(0), X_2(0)$, the $X_3(0), X_7(0), X_{15}(0), X_{16}(0), X_{18}(0), X_{19}(0), X_{20}(0), X_{21}(0), X_{22}(0), X_{23}(0), X_{24}(0), X_{26}(0), X_{27}(0), X_{28}(0), X_{29}(0), X_{30}(0), X_{31}(0), X_{32}(0), X_{33}(0), X_{34}(0), X_{35}(0), X_{36}(0), X_{38}(0), X_{39}(0), X_{40}(0), X_{41}(0), X_{42}(0), X_{43}(0), X_{44}(0), X_{46}(0), X_{47}(0), X_{48}(0), X_{49}(0), X_{50}(0), X_{51}(0)\}$;

$X_2(1) = \{X_4(0), X_5(0), X_{10}(0), X_{12}(0), X_{25}(0), X_{45}(0)\}$;

$X_3(1) = \{X_6(0)\}$;

$X_4(1) = \{X_8(0), X_9(0), X_{14}(0)\}$;

$X_5(1) = \{X_{11}(0)\}$;

$X_6(1) = \{X_{13}(0)\}$;

$X_7(1) = \{X_{37}(0), X_{17}(0)\}$.

Step 6. Perform secondary clustering for $X_1(1)$, $X_2(1)$, $X_3(1)$, $X_4(1)$, $X_5(1)$, $X_6(1)$ and $X_7(1)$. Repeat the operations from steps 2 to steps 5 above. The cohesion of $X_1(1)$ and $X_2(1)$ is *OAC* $(X_1(1), X_2(1)) < 65\%$. They cannot be combined. So $X_1(1)$ is denoted as $X_1(2)$ and $X_2(1)$ as $X_2(2)$. c = 2. The cohesion of $X_3(1)$ and $X_1(2)$ is *OAC* $(X_1(2), X_3(1)) < 65\%$. The cohesion of $X_3(1)$ and $X_2(2)$ is *OAC* $(X_2(2), X_3(1)) < 65\%$, c = 3.

Similarly, the second clustering result is still 7 classes, and it is the same as the first clustering result. The final result is shown as follows.

Class 1: $\{X_1(0), X_2(0), X_3(0), X_7(0), X_{15}(0), X_{16}(0), X_{18}(0), X_{19}(0), X_{20}(0), X_{21}(0), X_{22}(0), X_{23}(0), X_{24}(0), X_{26}(0), X_{27}(0), X_{28}(0), X_{31}(0), X_{32}(0), X_{33}(0), X_{34}(0), X_{36}(0), X_{38}(0), X_{39}(0), X_{40}(0), X_{42}(0), X_{43}(0), X_{44}(0), X_{46}(0), X_{47}(0), X_{48}(0), X_{49}(0), X_{50}(0), X_{51}(0)\}$;
Class 2: $\{X_4(0), X_5(0), X_{10}(0), X_{12}(0), X_{25}(0), X_{45}(0)\}$;
Class 3: $\{X_8(0), X_9(0), X_{14}(0)\}$;
Class 4: $\{X_{17}(0), X_{37}(0)\}$.
Class 5: $\{X_6(0)\}$;
Class 6: $\{X_{11}(0)\}$;
Class 7: $\{X_{13}(0)\}$;

The results show that applying the object-attribute space segmentation method to divide the high-dimensional sparse matrix of laws-bidding process units can divide 51 bidding process units into the above 7 classes. The objects of the same class have a high similarity. After pre-clustering, we found that the laws numbered 9, 10, 11, 12 and 13 were not used in the bidding process, so it is suggested to remove them from the regulations module in the association model. Table 4 shows the results of pre-clustering after removing the unused laws by the above algorithm. The table is constructed with 51 units of bidding process as rows and 9 laws related to bidding activities as columns. If a regulation is used in a bidding process, the corresponding table attribute of the two items is 1. Adjust the rows and columns to form the different shaded areas shown in the table. The bidding process units in the same shaded area have great similarity. In the process of compiling the bidding documents, a given law can be realized to automatically identify the bidding process unit in the same class as the bidding process unit applicable to this law. For the given bidding process unit, automatically match the law unit applicable to the bidding process unit in the same class.

5 Conclusion

Applying object—attribute space segmentation method in the process of bidding can help to divided the attribution of regulations and bidding process unit. This can help to form law unit and the bidding process unit respectively and to build laws—tender

Table 4 Object attribute preclustering result diagram

Laws Bidding process	14	5	3	4	1	2	6	8	7
17	1	1	0	0	0	0	0	0	0
37	1	1	0	0	0	0	0	0	0
1	1	1	1	1	0	1	1	0	1
2	1	1	1	1	1	1	1	1	1
3	0	1	1	1	0	1	1	1	1
7	0	1	1	1	0	1	1	1	1
15	0	1	0	1	1	1	1	0	1
16	1	1	0	1	1	1	1	1	1
18	1	1	1	1	0	1	1	1	1
19	1	1	0	1	0	1	1	1	1
20	0	1	1	1	0	1	1	1	1
21	1	1	0	1	0	1	0	1	1
22	1	1	0	0	0	1	1	1	1
23	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1
26	0	1	0	0	1	1	1	1	1
27	0	1	1	1	1	1	1	1	1
28	1	1	1	1	0	1	1	1	1
29	0	1	1	1	1	1	1	1	0
30	0	0	1	1	1	1	1	1	0
31	1	0	0	1	1	1	1	1	1
32	1	1	1	1	0	1	1	0	1
33	1	1	1	1	1	1	1	1	1

(continued)

related model. The implementation of structural description and storage of laws—bidding activity relationship can assist related person to inspect the tender documents and control the bidding risk.

Table 4 (continued)

34	0	1	1	1	0	1	1	1	1
35	0	1	0	1	1	1	1	0	1
36	1	1	0	1	1	1	1	1	1
38	1	1	1	1	0	1	1	1	1
39	1	1	0	1	0	1	1	1	1
40	0	1	1	1	0	1	1	1	1
41	1	1	0	1	0	1	0	1	1
42	1	1	0	0	0	1	1	1	1
43	1	1	1	1	1	1	1	1	1
44	1	1	1	1	1	1	1	1	1
46	0	1	0	0	1	1	1	1	1
47	0	1	1	1	1	1	1	1	1
48	1	1	1	1	0	1	1	1	1
49	0	1	1	1	1	1	1	1	0
50	0	0	1	1	1	1	1	1	0
51	1	0	0	1	1	1	1	1	1
4	0	0	0	0	0	1	1	0	0
5	0	0	0	0	0	1	1	1	0
10	0	0	0	0	0	0	1	1	0
12	0	0	0	0	0	1	1	1	0
25	0	0	0	0	1	1	1	1	0
45	0	0	0	0	1	1	1	1	0
8	0	0	0	0	0	0	0	1	1
9	0	1	0	0	0	0	0	1	1
14	0	0	0	0	0	0	1	1	1
6	0	0	0	0	0	0	1	0	0
13	0	0	0	0	0	1	0	0	0
11	0	0	1	0	0	0	0	0	0

References

1. S. Wu, X. Gao, M.Bastian, *Discovery of High-Dimensional Sparse Clustering Knowledge* (Metallurgical Industry Press, Beijing, 2003)
2. X. Gao, S. Wu,G. Wei, Spatial division of logistics attributes, in *2008 International Conference on Advanced Information Technology* (2008)
3. Q. Zhu, X. Gao, S. Wu, H. Chen, High-dimensional sparse data object—attribute space segmentation. Math. Pract. Underst. **41**(7), 184–189 (2011)
4. Y. Hu, A. Wang, Partitioning the object-attribute space for data mining based on the merger of object elements, in *International Conference on Logistics* (IEEE, 2015)
5. W. Wang, X. Li, X. Feng, S. Wang, A review of sparse subspace clustering. J. Automatics **41**(8), 1373–1384 (2015)
6. H. Xie, Z. Zhao, H. Lu, F. Wang, X. Liu, Sparse low-rank subspace clustering algorithm. J. Qingdao Univ. (Natural science edition) **30**(3), 64–68 (2017)

7. J. Wang, S. Wang, Z. Deng, Some problems in cluster analysis. Control Decis. **27**(3), 321–328 (2012)
8. J. Gao, Analysis and suggestions on common problems in compiling bidding documents. Sci. Technol. Inf. **16**(3), 218–220 (2016)
9. D. Wang, Improving the quality of bidding documents and ensuring the bidding effect. Mod. Econ. Inf. **14**, 133 (2018)

Research on Bidding Case Recommendation Algorithm Considering Bidding Features

Wenting Liang, Hao Liu, and Yaoyu Hu

Abstract In this paper a case recommendation matrix is proposed based on the content unit that can be used for reference in the bidding case. The case feature matrix is obtained based on the content recommendation algorithm, and the similarity between the new bidding case and the historical bidding case is calculated by using the cosine similarity algorithm. Based on the sequence of similarity calculation results, the similarity judgment threshold of the cases is determined, and historical bidding cases that are similar to the new bidding cases are obtained to form a set of recommended cases. In order to verify the feasibility and effectiveness of the proposed algorithm, an experimental analysis was performed using the algorithm. A set of cases that meet the requirements can be recommended by the recommendation algorithm proposed in this paper, which makes related business personnel more convenient in actual.

Keywords Recommendation algorithm · Case characteristics · Bidding

1 Introduction

In the bidding activities, in order to ensure the legitimacy of the bidding projects and maintain the order and related interests of the bidding, there are a large number of documents based on national laws and regulations, which contain national laws and regulations and internal regulations of the industry to ensure social development and stable demand. The standardized system has the characteristics of long-term unchanged, so the contents of the documents are roughly the same for similar bidding

W. Liang (✉) · H. Liu · Y. Hu
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: lwtxxgz@163.com

H. Liu
e-mail: ustbliutian@163.com

Y. Hu
e-mail: 13707537@chnenergy.com.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_46

cases. Based on this, the bidding case features of historical bidding cases are obtained through the content recommendation algorithm, a feature matrix is constructed, and the new bidding case feature matrix is formed by combining with the user-entered feature information of the new bidding case. And then use the cosine similarity algorithm to calculate the similarity between the characteristics of new tender cases and historical tender cases, search for similar historical bidding cases in the bidding case database, and recommend these cases and their bidding documents to business personnel.

The content-based recommendation algorithm does not rely on the user's past browsing history and evaluation. It mainly extracts the features of the user's selected objects, retrieves the objects with the greatest similarity to these features, and recommends them to the user [1]. The cosine similarity algorithm measures the similarity between the content represented by the two vectors by calculating the cosine value of the angle between the two vectors. The recommendation of the bidding case is to obtain the bidding case characteristics of the historical bidding case through the content recommendation algorithm, construct a feature matrix, combine the user-entered feature information of the new bidding case to form a new bidding case feature matrix, and use the cosine similarity algorithm to calculate the new bidding case The similarity between the characteristics and the characteristics of historical bidding cases can be used to obtain recommended targets.

Research on the algorithm of bidding case recommendation will help business personnel to quickly find historical bidding cases with similar characteristics when facing various bidding objects, obtain bidding files, and complete the preparation of bidding files by directly or indirectly using the content of historical bidding files jobs to help business personnel save working time and improve work efficiency.

2 Research Status

Since China promulgated the "Tendering and Bidding Law of the People's Republic of China" in 2000 [2], tendering is the most effective way to optimize the allocation of market economy and social resources in my country. However, according to statistics, 80% of the projects are not satisfactory after completion, which has a lot to do with the immature bidding field in my country.

The writing of bidding files is an important part of bidding activities. Because it involves a large number of national laws and regulations, there are often problems such as lack of content and imperfections. At this time, similar tendering cases that can provide reference value are particularly important.

In 2015, Zhang Haining built a platform recommendation system model that integrates user attributes based on the Chongqing Municipal Government's procurement agreement supply platform. By drawing on the Jaccard correlation coefficient, he proposed a new method of calculating user similarity to help purchasers effectively and quickly purchase, but lacks incremental data updates and cannot meet the requirements of future data volume growth [3]. In 2018, Cheng Peng researched the recommendation algorithm and similarity algorithm of the bidding files based on the Big Data platform for metalworking insurance. Using machine learning methods, the frequency-based recommendation algorithm and the random walk-based Personal Rank algorithm were used to implement An enterprise's personalized recommendation bidding announcement, but it can only achieve passive recommendation bidding cases, lacking flexibility and initiative [4]. In 2019, Li Siyang and Dai Zhengyu conducted research on the application of big data analysis in bidding projects, and proposed to create a database of bidding cases and use indexing functions such as clustering algorithms to select eligible bidding cases to achieve a reasonable selection of post-tender evaluation projects, but It is limited to the standardization of evaluation of bidding cases, and it is not used in the recommendation of bidding cases [5].

At present, there are still few domestic recommendation research on bidding cases. Only part of the bidding platforms have implemented passive recommendation bidding cases based on user positioning and behavior history, but lack the flexibility and the algorithm of recommend bidding cases based on case characteristics.

3 Recommendation Methods of Bidding

3.1 Core Concepts and Algorithm Construction Principles

3.1.1 Core Concept

- File Unit: The file unit is part of the file content that has thematic and logical independence in the overall content of the bidding file. It includes three types of file units: feature type, module type, and format type. These file unit parts can be used for reference or slightly modified without even modification can complete the preparation of some tender files in the new case.
- Bidding Case Feature Similarity: The feature matrix of the new tender case and the case feature matrix of all the tender cases in the bidding case database are calculated by using a cosine similarity algorithm.
- Case Similarity Judgment Threshold: Threshold determination requirements for judging whether new bidding cases and historical bidding cases constitute similar elements and become elements of the recommendation set.

3.1.2 Principle of Bidding Case Recommendation Algorithm or Constuction

- The recommendations are accurate. Fully understand the business process of the bidders, and use the powerful computing power of the computer to ensure the accuracy of the algorithm execution results, which meets the practical and quality requirements of the business staff.

- The recommendation process is efficient. Reduce the computational complexity of the recommendation process and improve the convenience of business staff operations.

3.2 Bidding Case Recommendation Process

Based on the construction principles of the case recommendation algorithm, the basic idea of determining the recommendation algorithm is as follows:

First, the feature matrix of historical bidding cases and new bidding cases is constructed. Secondly, the cosine similarity algorithm is used to calculate the similarity of the two matrices to determine the similarity and the threshold of case similarity judgment. A set of recommended bidding cases was finally determined (Fig. 1).

3.3 Content-Based Recommendation Algorithm

3.3.1 Bidding Case Characteristics

Obtaining the characteristics of bidding cases is the basis for recommending bidding cases. Based on the process of actual bidding activities and analysis of the key elements involved in the original bidding documents, it was found that the bidding activities mainly included 5 core elements, namely, project profile, qualification conditions, bid evaluation methods, suppliers and complaints/objections, which covered 17 analysis elements that can describe the characteristics of bidding activities (Fig. 2).

Because 17 analysis elements of bidding activities are directly used as case features, the excessive conditions may cause problems such as low similarity matching and high calculation complexity, which is not conducive to the generation of recommendation results. Therefore, in order to ensure that the data is complete when calculating the cosine similarity, The results are accurate, and it is necessary to appropriately delete the analysis elements of the bidding activities and remove the following three types of elements: ① the data is incomplete; ② the data cannot be used as the basis for the classification of the bidding activities to reflect the characteristics of the case; ③ the data is textual. The deletions are as follows:

The project object, project scope and project manager of the core elements of the project profile: ① The project object is a product obtained by modifying specific bidding object types or modifying a certain performance to meet customer needs; ② The project scope is specific to the specific Description of the products and results. ③ Project managers come from different enterprises, and the project activities they are responsible for may come from different fields. All three are diverse and are not suitable as a basis for the classification of bidding activities.

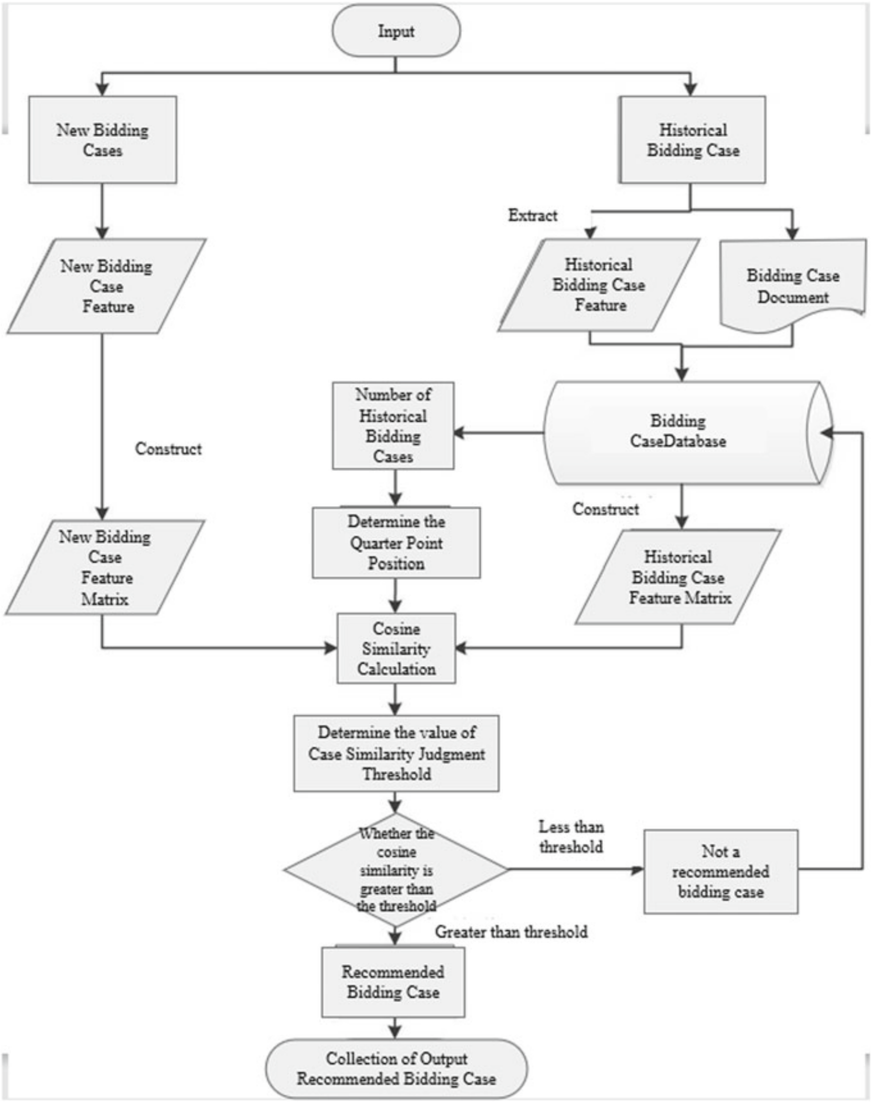

Input
New Bidding Cases
Historical Bidding Case
Extract
New Bidding Case Feature
Historical Bidding Case Feature
Bidding Case Document
Number of Historical Bidding Cases
Bidding CaseDatabase
Construct
Determine the Quarter Point Position
Construct
New Bidding Case Feature Matrix
Historical Bidding Case Feature Matrix
Cosine Similarity Calculation
Determine the value of Case Similarity Judgment Threshold
Whether the cosine similarity is greater than the threshold
Less than threshold
Not a recommended bidding case
Greater than threshold
Recommended Bidding Case
Collection of Output Recommended Bidding Case

Fig. 1 Tender case recommendation process

The qualification requirements for the core elements of the qualification conditions: belong to text data.

The scoring setting and scoring results of the core elements of the bid evaluation method: ① The scoring setting is built around the bidding objects, and the corresponding scoring is set according to certain properties and shapes of the bidding

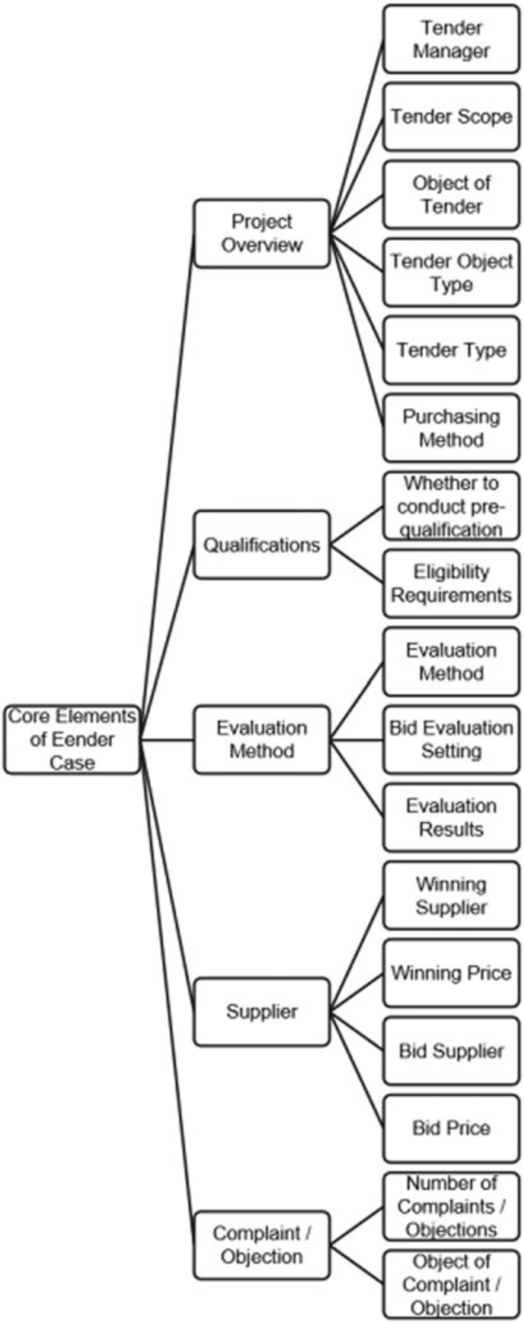


Fig. 2 Core elements and analysis elements of bidding activities

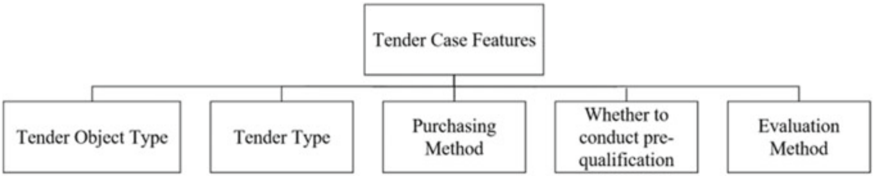


Fig. 3 Tender case feature model

objects required by the customers, which is not suitable as the basis for the classification of bidding activities. ② The scoring result is the result based on the evaluation method and scoring settings, and it is text data.

The core elements of the supplier: the failure of bidding will lead to partial loss of supplier data and cannot be filled.

Core elements of complaints/objections: Complaint/objection activities are caused by conflicts between parties in actual work, and they will exist in any field, which is not suitable as a basis for classifying bidding activities.

To sum up, after deletion, there are 5 analysis elements remaining in the 17 analysis elements, namely: project object type, bidding type, procurement method, whether to conduct qualification review and bid evaluation method, which constitute the characteristics of the bidding case and form the bidding case Feature model (Fig. 3). It is worth mentioning that the type of bidding object refers to the general classification of the articles allowed by the state in the actual bidding activities, which is limited and specific.

3.3.2 Tender Case Feature Matrix:

After obtaining the characteristics of the tender case, in order to calculate the cosine similarity, a matrix needs to be constructed. Therefore, the five analysis elements of the project object type, bidding type, procurement method, whether to perform qualification review and evaluation method are expressed as A_1, A_2, A_3, A_4, A_5, that is $T = [A_1, A_2, A_3, A_4, A_5]$, where T is the feature matrix of the bidding case.

Analyze the characteristics of the bidding cases in the bidding activities and find that the value range and data type of the characteristics of the bidding cases are shown in Table 1.

As the cosine similarity can only be calculated by numerical data, all data need to be converted to numeric data, and the data in the cosine similarity calculation result must have meaning, that is, the denominator is not 0. The conversion result is shown in Table 2.

The types of bidding objects (A_1): The types of bidding objects in this article are composed of the types contained in the experimental cases, including 57 values, which are assigned 1–57 respectively. After Min–Max standardization, the value range is narrowed [0, 1].

Table 1 Value range and data types of tender case characteristics

Tender case features	Ranges	Data type bidding
Object type (A_1)	57, heat exchangers, pumps, etc	Character
Character type tender type (A_2)	Goods; engineering; services	Character
Purchase method (A_3)	Open tender; invitation to tender	Character
Whether to undergo qualification examination (A_4)	Yes; no	Boolean
Boolean evaluation method (A_5)	Comprehensive bid evaluation method; lowest bid evaluation	Character

Table 2 Conversion results of the characteristics of the bidding case data types

Tender case features	Ranges	Data type bidding
Object type (A_1)	[0, 1]	Numeric
Character type tender type (A_2)	1: Goods; 2: Engineering; 3: Services	Numeric
Purchase method (A_3)	0: Open tender; 1: Invitation to tender	Numeric
Whether to undergo qualification examination (A_4)	0: Yes; 1: No	Numeric
Boolean evaluation method (A_5)	0: Comprehensive bid evaluation method; 1: Lowest bid evaluation	Numeric

Tender type (A_2): Includes 3 items of data, assigning goods, engineering, and services to 1, 2, and 3 respectively.

Procurement method (A_3): Including 2 items of data, the public bidding and invitation bidding are assigned 0 and 1, respectively.

Whether to conduct qualification review (A_4): Includes 2 items of data, and assigns 0 and 1 to the yes and no respectively.

Bid evaluation method (A_5): Includes 2 items of data, and assigns the comprehensive bid evaluation method and the lowest price bid evaluation method to 0 and 1, respectively.

For example, the bidding case type of a certain bidding case is "valve", the value is 0, the bidding type is goods, and the public bidding procurement method is used. Qualification review is required and the comprehensive bid evaluation method is used for bid evaluation. Tender Case Feature Matrix is [0, 1, 0, 1, 0].

3.4 Cosine Similarity Algorithm

The cosine similarity algorithm is an algorithm that measures the similarity between two vectors based on the result of measuring the cosine of the angle between the two vectors in the vector space. Its calculation formula is $\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\left|\vec{A}\right| \cdot \left|\vec{B}\right|} =$

$\frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n (A_i)^2} \times \sqrt{\sum_1^n (B_i)^2}}$, where $\vec{A}$ and $\vec{B}$ are vectors of two n dimensions, $\vec{A} = [A_1, A_2, A_3, \ldots, A_i, \ldots, A_n]$ and $\vec{B} = [B_1, B_2, B_3, \ldots, B_i, \ldots, B_n]$. The calculated cosine value, $\cos(\theta)$, ranges from $[-1, 1]$, that is, the closer the cosine value is to 1, the closer the angle between the two vectors is to 0°, the higher the similarity, the lower the dissimilarity. Conversely, the closer the cosine value is to -1, the lower the similarity and the higher the dissimilarity. When the cosine value is equal to 0, it means that the two vectors are independent.

Because there are 5 bidding case features in the bidding case recommendation algorithm, the vector corresponding to the case feature matrix of the *i-th* bidding case *Ci* in the bidding case database is set as $\vec{A_i} = [A_{i1}, A_{i2}, A_{i3}, A_{i4}, A_{i5}]$, The vector corresponding to the feature matrix of the *j-th* new bidding case N_j is set as $\vec{B_j} = [B_{j1}, B_{j2}, B_{j3}, B_{j4}, B_{j5}]$. Because the values are not negative, the included angle between the vectors corresponding to the matrix is between 0° and 90°, so $\cos(\theta) \in [0, 1]$, and the larger $\cos(\theta)$, the smaller the included angle between the vector $\vec{A_i}$ and the vector $\vec{B_j}$, indicating that the Bidding Case Feature Similarity between the *i-th* bidding case C_i in the bidding case database and the *j-th* new bidding case N_j is higher. The cosine similarity algorithm formula in the bidding case recommendation algorithm is obtained as follows:

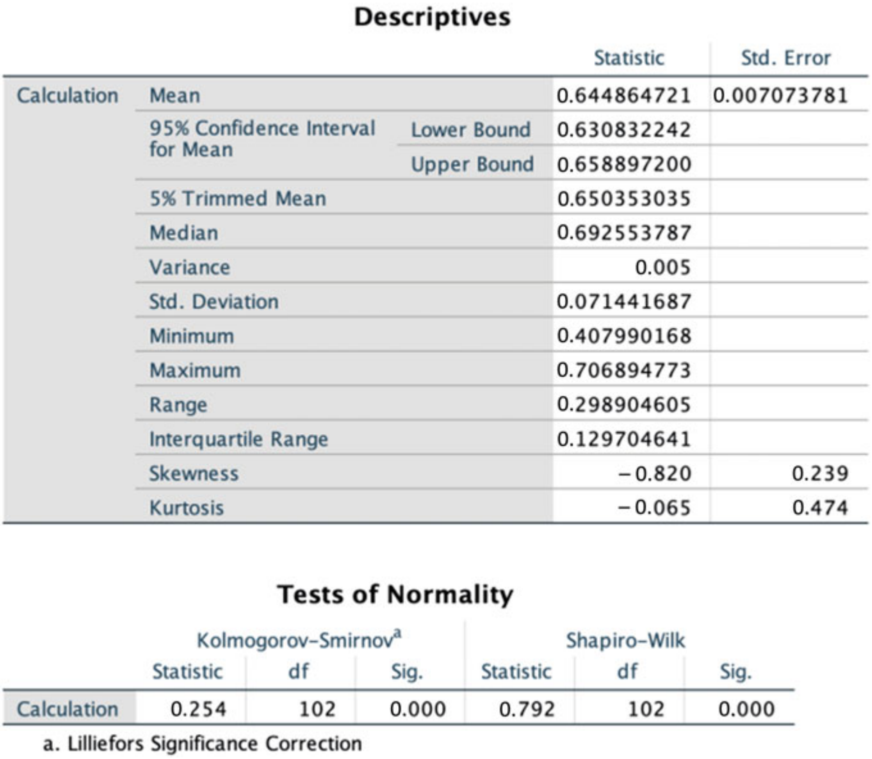
$$\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\left|\vec{A}\right| \cdot \left|\vec{B}\right|} = \frac{\sum_1^5 \left(A_{it} \times B_{jt}\right)}{\sqrt{\sum_1^5 (A_{it})^2} \times \sqrt{\sum_1^5 \left(B_{jt}\right)^2}},$$

where $t = 1, 2, 3, 4, 5$ and t represents 5 attributes of the characteristics of the bidding case.

3.5 Case Similarity Judgment Threshold

Set the Case Similarity Judgment Threshold to $\beta \in [0, 1]$, and when $\cos(\theta) \in [\beta, 1]$, explain that the two vectors meet the required similarity, and determine that the corresponding historical bidding case is similar to the new bidding case. This historical bidding case becomes an element in the set of recommended bidding cases. When $\cos(\theta) = [0, \beta]$, the cases are not determined to be similar, and these cases were not in the set of recommended bidding cases.

The bidding case recommendation algorithm is used to calculate different practice case libraries, and the obtained cosine similarity calculation results are sorted in descending order. Using SPSS software to analyze many similarity calculation result sequences, the results are consistent, and the data description and normal distribution test of one similarity calculation result sequence are shown in Fig. 4. In the normality test, *sig.* < 0.05, that is, the significance is less than 0.05, does not belong to a normal



Descriptives

			Statistic	Std. Error
Calculation	Mean		0.644864721	0.007073781
	95% Confidence Interval for Mean	Lower Bound	0.630832242	
		Upper Bound	0.658897200	
	5% Trimmed Mean		0.650353035	
	Median		0.692553787	
	Variance		0.005	
	Std. Deviation		0.071441687	
	Minimum		0.407990168	
	Maximum		0.706894773	
	Range		0.298904605	
	Interquartile Range		0.129704641	
	Skewness		−0.820	0.239
	Kurtosis		−0.065	0.474

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Calculation	0.254	102	0.000	0.792	102	0.000

a. Lilliefors Significance Correction

Fig. 4 Data description and normal distribution test of a similarity calculation result

distribution, and the Skewness is less than 0, indicating that the similarity calculation result is a negative skewed distribution in the [0, 1] interval.

Corresponding to the degree of similarity is the degree of difference, and *the degree of similarity = 1 − the degree of difference*. When the degree of similarity is sorted from large to small, the degree of difference is sorted from small to large. Because the interquartile range is the best indicator to describe the variability of a set of skewed data, the case similarity threshold is initially set to the case corresponding to the three quartiles of the sequence of the degree of difference calculation results arranged from small to large Similarity [6]. On the whole, the difference between the calculation results of similarity is not large, but by calculating the standard deviation, it is found that the standard deviation of the last 25% of the sequence of similarity calculation results is generally larger than the standard deviation of other intervals, indicating that the degree of variability of the data is larger. Similarly, the variability of the calculation results of the difference in this part of the data is also large. Therefore, the difference calculation result of the third quartile is selected, that is, the difference degree α of the Q3 position of the difference calculation sequence. Then the corresponding similarity is $\beta = 1 - \alpha$, and the historical bidding cases whose similarity calculation result is greater than or equal to β constitute the recommended case set.

Assume that there are n historical bidding cases in the bidding case database C_i, calculate the cosine similarity and difference, and sort the differences in ascending order, and let $m = 3(n + 1)/4$, where m represents the position of the quartile, and when m is an integer, the value of β is the similarity corresponding to the *m-th* term in the sequence. If m is a fraction and the decimal part is γ, then $\alpha = (1 - \gamma) \times$ *the value of the [m]-th term* $+ \gamma \times$ *the value of the* $([m] + 1)$-*th term.*

4 Experiment and Result

The bidding activities are distributed in various fields, and the case recommendation of the bidding documents in this article also applies to the reference document units in the bidding activities in various fields. The relevant data of the bidding cases involved in this experimental project come from the Shenning Group's 4 million tons a year coal-to-oil project [7]. After data selection and processing, a total of 152 bidding cases and 5 case bidding case characteristics were numerically expressed. As the basic data for conducting experiments.

In order to divide the 152 cases into two parts without being affected by human factors, as the historical bidding case database and the new bidding case database in the experiment, this study uses the random selection function in Excel to conduct a random selection from 152 bidding cases and 152 results were selected. After the operations of deduplication and deletion of blanks, a total of 99 bidding cases constituted a historical bidding case database, and the remaining 53 constituted new bidding cases.

According to the bidding case recommendation process, randomly select C148 as the new bidding case, obtain the case characteristics of the new bidding case and the case characteristics of all historical bidding cases in the bidding case database, and convert the case characteristics into numerical data to form a case characteristic matrix. The calculation formula of the cosine similarity algorithm is calculated to obtain the cosine similarity. Since $m = 3 \times (99 + 1) \div 4 = 75$, m is an integer, and β is the similarity corresponding to the *m-th* item. After sorting the similarity in descending order, we know that $\beta = 0.565646518$. The historical bidding cases with a similarity greater than or equal to β, that is, cases with descending numbers from 1 to 75, constitute the recommended bidding case set and are recommended to business personnel, including the bidding documents of each case and other materials, to help them prepare the bidding documents for new bidding cases, and other bidding cases are waiting for the next recommendation calculation in the bidding case database.

5 Summary and Outlook

This article obtains a case feature matrix is obtained based on the content recommendation algorithm, the similarity between the new bidding case and the historical bidding case is calculated by using the cosine similarity, the recommended case according to the case similarity judgment threshold is determined, and the recommendation of bidding documents is implemented to business personnel to make it work more convenient in practice, which is the goal of this study. In addition, after the new bidding case is realized in actual work, the file unit and case characteristics can be directly included in the bidding case database, which becomes a historical bidding case and has flexibility.

Because the experimental data in this article is from the same project, the research scope is limited. Besides, the calculation of cosine similarity is affected by the extraction of case features and numerical processing, and than deviation is retained. In addition, the value of the Case Similarity Judgment Threshold depends on the number of historical bidding cases and the calculation results of similarity, and the lack of accuracy is existed.

References

1. L. Luo, Leadership decision-making auxiliary information system based on content recommendation algorithm and cosine similarity algorithm. J. Guangxi Acad. Sci. **34**(02), 143–150 (2018)
2. L. Zhang, Bidding evaluation of engineering projects based on improved analytic hierarchy process method research. Southwest University of Science and Technology (2019)
3. H.N. Zhang, Research and application of cargo platform recommendation system. Chongqing University of Science and Technology (2015)
4. P. Cheng, Research on personalized recommendation and competitiveness prediction based on metalworking insurance bidding data. Wenzhou University (2018)
5. S.Y. Li, Z. Dai, Research on the application of big data analysis in post-evaluation of bidding projects. Tendering Procurement Manag. (02), 65–67 (2019)
6. D.J. Qi, M. Chen, On the calculation of quartiles. Available online: https://wenku.baidu.com/view/2e33fd62caaedd3383c4d3b3.html (accessed on 18 February 2020). (In Chinese)
7. S.P. Liu, Shenhua Ningxia coal industry group 4 million tons/year coal indirect liquefaction project safety facility design special review meeting. Refinery Technol. Eng. **45**(08), 56 (2015)

Pricing and Versioning Strategy for Information Products with No-Free-Disposal: The Role of Consumers' Expectations Formation

Danqin Yang and Juan Bai

Abstract In this paper, we consider a monopolist providing two versions of products with no-free-disposal (NFD) in the information markets. Based on the maximization of profit and welfare, we get the optimal solutions in the case of exogenous quality and endogenous quality respectively depending on whether consumers' expectation of market share is affected by price or not. Through analysis and comparison, we find that the optimal versioning strategy of the monopolist is a single high-quality version based the welfare maximizing. Compared to the case where consumers' expectation is not affected by the market price, the monopolist can obtain more expected market share and profits by regulating the market price in the case where consumers' expectation is affected by the market price.

Keywords Versioning · Information goods · Pricing · Expectations · Free disposal · Network effects

1 Introduction

With the rapid development of network and information technology, the versioning strategy has become a common practice for the vertical differentiation of information products in information markets. However, many previous studies have considered that the two versions of information products as the sales strategies of suppliers [1–5]. The nature of information products and services is participatory. Participation requires consumers to pay a certain amount of time and learning costs. Therefore, information products and services have the feature of no free disposal. This means that consumers will be less willing to buy more at the same price. In this paper, we consider that the utility function of consumers is nonmonotonic in quality to reflect

D. Yang (✉) · J. Bai
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China
e-mail: yangdanqin@163.com

J. Bai
e-mail: 18356359844@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_47

this kind of unsatisfaction. In other words, if the functions of software products are higher than the functions required by users, their willingness to pay will be affected by the higher cost of using resources. This kind of resource use cost may be due to the higher version of information products often need to occupy more storage space. Take spreadsheet software products as an example. A statistician who wants to complete the task of demand forecasting may need Microsoft Excel 2010 to provide multiple functions. Compared with statisticians, a student only needs a single function of Microsoft Excel 2010 to complete the homework. Thus, when installing Microsoft Excel 2010 and other software, the student must consider the impact of additional learning costs and memory space occupied by the software, which undoubtedly reduces his willingness to buy software.

To the best of our knowledge, previous studies rarely consider the influence of the formation mechanism of consumers' expectation of firms' different products market share. In this paper, based on the profit maximization and welfare maximization, we get the optimal solutions in the two cases respectively. Specifically, we emphasize the influence of consumers' expectations on the pricing decision-making of the monopolists.

The rest of this paper is organized as follows. In the next section, we review the most relevant literature. Model settings are described in Sect. 3. In section 4, based on welfare maximization, we compare and analyze the optimal solution depending on whether the price is affected by expectation. Finally, we conclude the paper and discuss the future research.

2 Literature Review

Our work is closely related to versioning strategy, no-free-disposal, pricing of information products, consumer's expectation.

Jing [1] studied the influence of network externalities on versioning strategy. They found a multi-product monopolist offers two versions of distinct qualities. Li et al. [2, 3] studied a bilevel programming model to represent the task for optimizing the strategy of versioning an information product. Liu et al. [4] provided a quantitative analysis of selling information products, aiming to determine the best sale channel and versioning strategy in the presence of network externality. Shivendu and Zhang [6] studied versioning strategies, while they focused on the "inconvenience" or uselessness that users will encounter when the function of the software is lower than the function required to complete the task.

Chellappa and Shivendu [7] examined vendor strategies in a market where consumers have heterogeneous concerns about privacy based online personalization services with a "no free disposal" property. Chellappa and Mehra [8] studied the influence of the marginal cost of the monopolist and the use cost of the consumer on the version control strategy for the goods without free disposal. Jing [9] examined how customer learning investment choices interact with quality and price competition in a duopoly. Griva and Vettas [10] studied the role of consumer expectations in duopoly market where products differ horizontally and vertically.

Our paper is closely related to [8] and [10]. Chellappa and Mehra [8] mainly emphasizes the impacts of monopolist's marginal cost and consumers' usage costs on versioning strategy. Different from [8], we consider the influence of consumer expectation in the market of information products with a monopolist. Griva and Vettas [10] emphasized the influence of consumers' expectations on duopoly market share, we further consider whether consumers' expectation of market share is affected by price, which has an impact on monopolist pricing decisions.

3 Model Settings

In this paper, we consider the market in which a monopolist provides two versions of information products (i.e. the products with high-quality and the products with low-quality). When the network externality exists, a consumer purchases the product with a network size N at a price p, and he/she can obtain the net utility [4]:

$$(\theta + \gamma N)q - \lambda q^2 - p \tag{1}$$

where θq is the product's inherent "network-independent" or standalone value, and γNq is the "network-generated" value derived from network externality. Observe that the utility function is nonmonotonic in quality, i.e., the utility is first increasing in quality, and up to a point (at which utility is maximized) then it decreases. In other words, higher quality can actually make the consumer worse off. For example, high-quality versions of information products often require higher memory space. Consumers may consider the "no free disposal" (NFD) property when purchasing high-level information products. And the property of information products is expressed in this paper as the cost of resources for consumers, namely $\lambda q^2 (\lambda > 0)$.

The definitions of some variables and parameters are listed in the Table 1.

3.1 Based on Monopolist's Profit Maximization (Exogenous Quality)

When a monopolist provides two types of differentiated products in the market, the net utility of the consumer who buys the low-quality product is

$$U_l = \left(\theta + \gamma N^e\right)q_l - \lambda q_l^2 - p_l \tag{2}$$

Table 1 Notation

Symbol	Definition
$q(q_h, q_l)$	The product quality level (the quality of the high-end version, the quality of the low-end version)
θ	The degree of consumers' preference for quality, $\theta \sim U(0, 1)$
γ	Marginal network externality value, $\gamma \in [0, 1]$
$N(N_l, N_h, N^e)$	The total market demand (the demand of the high-end version, the demand of the low-end version and the expected total market demand), $N(N_l, N_h) \in [0, 1]$
$p(p_h, p_l)$	The price decided by the firms (the price of the high-end version, the price of the low-end version)
c	A fixed, quality-dependent cost of creating the highest quality
λ	The resource-cost coefficient, $\lambda > 0$

And the net utility of the consumer who buys the high-quality product is

$$U_h = \left(\theta + \gamma N^e\right)q_h - \lambda q_h^2 - p_h \tag{3}$$

The profit of the firm providing two types of products of different quality is:

$$\prod = p_l N_l + p_h N_h \tag{4}$$

Here, since the quality of information products is exogenous, the R&D costs related to product quality can be regarded as sunk costs, i.e.,fixed costs.

As is known to all, in addition to comparing the prices set by the firms, consumers may choose a product with higher quality, because they expect that a larger number of other users will choose the same product (and hence they expect its value to increase via the network effect). The latter can be expressed as the consumers' expectation for market share of different products in the monopolized information market. Throughout the analysis, consumers' expectations about firms' different products market shares are required to be fulfilled in equilibrium, to be "rational". However, we distinguish between two scenarios. First, we examine the case where these expectations cannot be influenced by the prices set by the firms. In this case, expectations are formed before the prices are set (or equivalently, even if expectations are formed after some prices are set, that does not matter because firms do not commit to these prices). In a two stage game, first firms announce their prices, subsequently consumers, based on their expectations and the announced prices maximize their utility by selecting a unit of a product. We are looking for a subgame perfect Nash equilibrium and solving backwards. First the demand function is determined and then, given this demand, firms' profits have to be maximized with respect to the prices. Then we solve the model assuming rational (in other words, fulfilled in equilibrium) expectations. Second, we examine the case where consumers' expectations can be influenced by the prices set by the firms; this should be the case when firms commit to the prices. In this scenario, firms announce their prices, knowing that these announcements will influence consumers' expectations about each firm's market share. Then, consumers form their expectations, taking as given the prices that have been announced and choose which product to purchase. Since firms can influence consumers' expectations via their prices, these prices should be used when deriving the expected market shares (which, of course, will be equal in equilibrium to the actual market shares).

In this paper we emphasize the role of consumers' expectations formation when there is network externality on the pricing decision-making of the monopolist.

3.1.1 Optimal with Expectations not Influenced by Prices

In this section, as [10] studied the role of consumers' expectations formation in duopoly market, we analyzed the situation where consumers' expectations of market share (i.e., network effects) are not affected by published prices at optimal.

From Eqs. (1)–(4), we derive the following proposition.

Proposition 1 *Assume consumers' expectations are not influenced by prices. If* $0 < \lambda q_l < \lambda q_h < 1, 0 < \gamma < 1$ *or* $0 < \lambda q_l < 1 < \lambda q_h < 2$, $\frac{2(\lambda q_h - 1)}{\lambda(q_h - q_l)} < \gamma < 1$, *then the monopolist offers two versions of products and we can get the optimal prices, demands and profit of the firm with profit maximization as follows:*

$$
\begin{aligned}
p_h^{PNE^*} &= \frac{q_h\left[2(1-\lambda q_h)+\lambda\gamma(q_h-q_l)\right]}{2(2-\gamma)}\\
p_l^{PNE^*} &= \frac{q_l(1-\lambda q_l)}{2-\gamma}\\
N_l^{PNE^*} &= \frac{\lambda q_h}{2}\\
N_h^{PNE^*} &= -\frac{2\lambda(q_l+q_h)-\lambda\gamma q_h-2}{2(2-\gamma)}\\
N^{PNE^*} &= \frac{1-\lambda q_l}{2-\gamma}\\
\pi^{PNE^*} &= \frac{q_h\left(4\lambda^2 q_l^2(\gamma-1)-\lambda^2 q_l q_h\left(\gamma^2-4\right)+\lambda^2 q_h^2(\gamma-2)^2+4\lambda\gamma(q_h-q_l)-8\lambda q_h+4\right)}{4(2-\gamma)^2}
\end{aligned}
$$

The superscript PNE here indicates expectations not influenced by prices based on profits' maximization.

Proofs of all propositions and corollaries are given in the Appendix.

3.1.2 Optimal with Expectations Influenced by Prices

Now, we study how prices and market share are formed when companies can manipulate consumer expectations. Companies announce prices, and consumers consider published prices when forming expectations. In this case, the monopolist will publish the price because they know that these announcements will affect consumers' expectations of monopolist's market share. Thus, consumers form their expectations and choose which product to buy based on the published price.

From Eqs. (1)–(4), we derive the following proposition.

Proposition 2 *Assume consumers' expectations are influenced by prices. If*

$$\begin{cases} 0 < \lambda < \frac{3\gamma q_l - 2q_l + \gamma q_h + \gamma^2 q_h - \gamma^2 q_l}{\gamma\left(q_h^2 - q_l^2\right) + 2q_l^2}, \\ 4q_l(1-\gamma) + \gamma^2(q_l - q_h) > 0, \\ 1 - \lambda q_h < 0, \\ \frac{2q_l(\lambda q_h - 1)}{\lambda(q_h - q_l)} < \gamma < 1. \end{cases}$$

or

$$\begin{cases} 0 < \lambda < \frac{3\gamma q_l - 2q_l + \gamma q_h + \gamma^2 q_h - \gamma^2 q_l}{\gamma\left(q_h^2 - q_l^2\right) + 2q_l^2}, \\ 4q_l(1-\gamma) + \gamma^2(q_l - q_h) > 0, \\ 1 - \lambda q_h > 0, \\ 0 < \gamma < \frac{2q_l(1 - \lambda q_l)}{(q_h - q_l)[\lambda(q_l + q_h) - 1]}, \\ 1 - \lambda(q_l + q_h) < 0. \end{cases}$$

then the monopolist offers two versions of products and we can get the following optimal prices, demands and profit of the firm with profit maximization:

$$p_h^{PE^*} = \frac{q_l q_h (1-\gamma)\left[2(1-\lambda q_h) + \lambda\gamma(q_h - q_l)\right]}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

$$p_l^{PE^*} = \frac{q_l\left(\lambda\gamma\left(q_l^2 + q_h^2\right) + \lambda\gamma^2 q_h(q_l - q_h) - 2\lambda q_l^2 - \gamma(q_l + q_h) + 2q_l\right)}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

$$N_l^{PE^*} = \frac{q_h(\gamma + 2\lambda q_l(1-\gamma) - \lambda\gamma q_h)}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

$$N_h^{PE^*} = \frac{-(\gamma - 2 + 2\lambda(q_l + q_h) - \lambda\gamma q_l - 2\lambda\gamma q_h)q_l}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

$$N^{PE^*} = \frac{2q_l(1 - \lambda q_l) + \gamma(q_h - q_l)[1 - \lambda(q_l + q_h)]}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

$$\pi^{PE^*} = \frac{q_l q_h\left(1 + \lambda^2\left(q_h^2 - q_l^2 + q_l q_h\right) - 2\lambda q_h + \lambda^2\gamma\left(q_l^2 - q_h^2\right) - \lambda(q_l - q_h)\right)}{4q_l(1-\gamma) + \gamma^2(q_l - q_h)}$$

The superscript PE here indicates expectations influenced by prices based on profits' maximization.

Considering that the quality is exogenous given in the context of maximizing the profit of the monopolist, we only give the optimal pricing decision of the monopolist here as a reference.

3.2 Based on Monopolist's Welfare Maximizing (Endogenous Quality)

The schedule of the model is as follows: the seller invests in R&D to produce the highest quality products. Then he can create other versions of products with lower quality by reducing some functions or removing some configurations of higher version products and set prices accordingly. There is no cost for the supplier to create a version of reduced quality (zero version conversion cost), and it will not bring any additional cost (zero marginal cost) to the service consumer, which is determined by the characteristics of information products. To give priority to determine this highest quality level, the supplier then considers his next stage decision of versioning and corresponding prices and employs backward induction. In this paper, we only consider two versions of products provided by a monopolist. Please note that the solution process of this problem is consistent with the profit maximization, but the seller will extract all the rest.

We suppose that q_h is the high-quality level generated by the seller, and $q_h(\theta)$ is the quality enjoyed by type of θ consumers. Please note that our utility function is nonmonotonic concave. That is, every consumer has a saturation point at which he can get the maximum benefit from consumption. This utility function is maximized at

$$q_h^*(\theta) = \arg\max_q \big[\big(\theta + \gamma N^e\big) q_h - \lambda q_h^2 - p_h\big] \tag{5}$$

Here, $q_h^*(\theta)$ is the quality of high version products based on consumers' preferences. Given the high quality product, the monopolist extracts the full surplus of the consumer which can be obtained from

$$U(\theta) = \big(\theta + \gamma N^e\big) q_h^*(\theta) - \lambda q_h^*(\theta)^2 - p_h = 0 \tag{6}$$

The participation constraint

$$U = (\theta + \gamma N) q - \lambda q^2 - p \geq 0$$

From Eqs. (5) and (6), we derive the following proposition.

Proposition 3 *Assume the monopolist pursues the welfare maximizing. If* $p_h = \lambda q_h^2$, $p_l = \lambda q_l(2q_h - q_l)$,*then the monopolist extract largest surplus from high type consumer.*

In the complete information case, the seller knows the two types of consumers in the market of information products; therefore, given the high quality product, the supplier prefers to extract the maximum surplus of the consumer who prefers the high-quality version.

3.2.1 Optimal with Expectations Not Influenced by Prices

The profit of the firm providing two types of products of differentiated quality is as follows:

$$\prod = p_l N_l + p_h N_h - cq_h^2 \tag{7}$$

where $p_h = \lambda q_h^2$, $p_l = \lambda q_l(2q_h - q_l)$.

Here, because product' high quality is endogenous, we show the product's R&D costs as a quadratic function of quality.

From Eq. (7), we derive the following proposition.

Proposition 4 *Assume consumers' expectations are not influenced by prices. If* $c < \lambda < 2c, 0 < \gamma < \frac{2(\lambda - c)}{\lambda}$ or $\lambda > 2c, 0 < \gamma < 1$, *then the monopolist offers the single high version and the corresponding optimal price, demand, quality and profit of the firm with welfare maximization is as follows*

$$\begin{aligned} N_h^{WNE^*} &= \frac{\lambda + 2c}{\lambda(3-\gamma)} \\ q_h^{WNE^*} &= \frac{\lambda - c + c\gamma}{\lambda^2(3-\gamma)} \\ p_h^{WNE^*} &= \frac{(\lambda - c + c\gamma)^2}{\lambda^3(3-\gamma)^2} \\ \pi^{WNE^*} &= \frac{(\lambda - c + c\gamma)^3}{\lambda^4(3-\gamma)^3} \end{aligned}$$

The superscript WNE here indicates expectations not influenced by prices based on welfare maximizing.

From Proposition 4, we can draw the following conclusion: in order to extract the largest surplus from consumers, the monopolist can only provide single type of high quality products to the market since the demand of the type of low quality products is zero. In this case, the target market of the monopolist is the consumers with high-quality version preference, and the maximum surplus extracted by monopolist from the consumers with high-quality version preference is enough to make up for the loss caused by the losing of market share of consumers with low-quality version preference.

3.2.2 Optimal with Expectations Influenced by Prices

From Eq. (7), we derive the following proposition.

Proposition 5 *Assume consumers' expectations are influenced by prices. If* $\lambda > c, 0 < \gamma < 1$ or $\lambda < \frac{2c}{3}, \frac{2c-2\lambda}{3\lambda-2c} < \gamma < 1$, *then the monopolist offers the single high version and we can get the following optimal price, demand, quality and profit of the firm with welfare maximization:*

$$
\begin{aligned}
q_h^{WE^*} &= \frac{\lambda - c + c\gamma}{3\lambda^2} \\
N_h^{WE^*} &= \frac{\lambda + 2c - 2c\gamma}{3\lambda(1-\gamma)} \\
p_h^{WE^*} &= \frac{(\lambda - c + c\gamma)^2}{9\lambda^3} \\
\pi^{WE^*} &= \frac{(\lambda - c + c\lambda)^3}{27\lambda^4(1-\gamma)}
\end{aligned}
$$

The superscript WE here indicates expectations influenced by prices based on welfare maximizing.

From Propositions 4 and 5, we derive the following corollary.

Corollary 1 *Assume the monopolist pursues the maximization of welfare, no matter whether the consumers' expectation of market share is affected by the market price or not, the optimal price, quality, demand of the high version and the optimal profit of the firm increase with the network externalities.*

In this case, although monopolists only provide a single version of high-quality products in order to extract the total surplus of consumers, network externalities have a positive impact and further increase the market share of the monopolist.

4 Comparative Analysis

On the basis of Sect. 3, we can further analyze and discuss the situation of monopolists pursuing welfare maximization.

From Proposition 4 and 5, we derive the following corollary.

Corollary 2 *Assume the monopolist pursues the maximization of welfare, when the consumers' expectation is not affected by the market price, the optimal price and quality of the high version are higher than that when the consumers' expectation is affected by the market price, but the optimal market share and the optimal profit of the monopolist are lower than that when consumers' expectation is affected by the market price.*

In the case of the maximization of welfare (quality is endogenous), when consumers' expectations are not affected by market prices, it corresponds to the situation that consumers think that companies cannot guarantee their published prices, or consumers' expectations are formed before prices are observed. In this way, for a monopolist, he will not regard price as a tool to manipulate market share. When consumers' expectation is affected by market price, the monopolist can obtain the expected market share by regulating market price, i.e., monopolist can obtain a larger market share through lowering products' quality, so as to obtain a higher profit.

Next, based on Corollary 2, we make a simple numerical analysis of the situation that consumers' expectations are not affected by prices. Here, in order to simplify the analysis, we only consider the case of $\gamma = 0.5$.

We can intuitively see how $p_h^{WE^*}$, $q_h^{WE^*}$, $N_h^{WE^*}$ and π^{WE^*} change with λ or c according to Fig. 1.

According to Fig. 1, when the resource cost coefficient of consumers λ and the product R&D cost coefficient of monopolists c are low, the price and quality of information products are relatively high, and the market share of the monopolist is relatively low. In this situation, the higher market pricing leads to the loss of market share, and the profits of the monopolist is also reduced. In this case, because the target market of the monopolist is the consumers with high version preference, the monopolist provides a single version of high-quality information products. On the other hand, the R&D cost coefficient of the monopolist is low, and the monopolist will

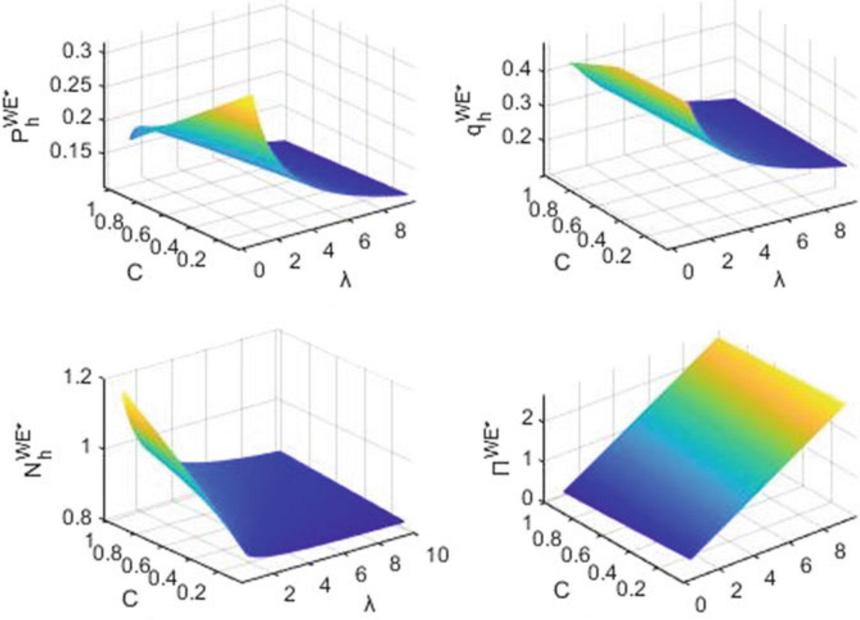


Fig. 1 Optimal price, quality, demand and profit versus λ and c

be committed to providing higher quality products, leading to higher product pricing. At the same time, the lower resource cost coefficient of consumers can't make up for the loss caused by the increase of product price which leads to a reduction in market demand.

We can also see when the resource cost coefficient of consumers is largest, the profit of monopolists is highest. In this case, the monopolist will consider providing lower quality products with lower market pricing to obtain more market share.

5 Conclusion

In this paper, we can obtain the following results. First of all, when monopolist pursues profit maximization, we give the optimal pricing decision of monopolist as a reference under exogenous quality considering the expectation is influenced by market price or not.

Secondly, when the monopolist considers the maximization of welfare, the production plan of the supplier will give priority to the investment in the production of high-level information products, which is determined by the characteristics of the information products, i.e., zero transformation cost. Regardless of whether consumers' expectation of market share is affected by market price or not, we found the optimal versioning strategy changed from providing two versions of information-based products to only providing high-quality version of informational products. Part of the reason may be that the maximum surplus extracted by monopolists from high-quality version preference consumers is enough to make up for the loss caused by the losing of the market share of consumers with low-quality version preference.

At last, through simple analysis and comparison, we can draw the conclusion that when consumers' expectation is affected by market price, the monopolist can obtain the expected market share by regulating market price. That is, monopolist can obtain a larger market share through lowering products' quality, so as to obtain a higher profit.

Our paper has several limitations. First, the supply chain we considered is under complete information. Second, in order to simplify the problem, we consider that the cost of consumer resource use is homogeneous. Third, the information market we consider is full coverage. Future work can consider incomplete information, the cost of consumer resource use is heterogeneous, and not completely covered information market.

Acknowledgements This work is supported by National Natural Science Foundation of China under grant 71771123.

Appendix: Proofs

Proof of Proposition 1 We assume that θ_1 indicates that there is no difference between consumers' purchase and non-purchase of low-quality products, and θ_2 indicates that there is no difference between consumers' purchase of high-quality products and low-quality products.

$$\left(\theta + \gamma N^e\right)q_l - \lambda q_l^2 - p_l = 0 \tag{8}$$

$$\left(\theta + \gamma N^e\right)q_h - \lambda q_h^2 - p_h = \left(\theta + \gamma N^e\right)q_l - \lambda q_l^2 - p_l \tag{9}$$

Therefore, we have

$$\begin{cases} \theta_1 = \dfrac{p_l}{q_l} + \lambda q_l - \gamma N^e \\ \theta_2 = \dfrac{p_h - p_l}{q_h - q_l} + \lambda(q_h + q_l) - \gamma N^e \end{cases} \tag{10}$$

The demand for high and low quality products is

$$\begin{cases} N_l = \theta_2 - \theta_1 = \dfrac{p_h - p_l}{q_h - q_l} - \dfrac{p_l}{q_l} + \lambda q_h \\ N_h = 1 - \theta_2 = 1 - \dfrac{p_h - p_l}{q_h - q_l} - \lambda(q_h + q_l) + \gamma N^e \\ N = 1 - \theta_1 = 1 - \dfrac{p_l}{q_l} - \lambda q_l + \gamma N^e \end{cases} \tag{11}$$

Thus, the profit of the firm providing two products of differentiated quality is as follows:

$$\begin{aligned} \max_{p_l, p_h} \prod^{PNE} &= p_l^{PNE} N_l^{PNE} + p_h^{PNE} N_2^{PNE} \\ &= p_l^{PNE}\left(\frac{p_h^{PNE} - p_l^{PNE}}{q_h - q_l} - \frac{p_l^{PNE}}{q_l} + \lambda q_l\right) \\ &\quad + p_h^{PNE}\left(1 - \frac{p_h^{PNE} - p_l^{PNE}}{q_h - q_l} - \lambda(q_h + q_l) + \gamma N^e\right) \end{aligned} \tag{12}$$

Here, the Hessian matrix is as follows

$$H_1 = \begin{bmatrix} -\frac{2q_h}{(q_h - q_l)q_l}, & \frac{2}{q_h - q_l} \\ \frac{2}{q_h - q_l}, & -\frac{2}{q_h - q_l} \end{bmatrix}$$

From $-\frac{2q_h}{(q_h-q_l)q_l} < 0$, $\frac{4}{q_l(q_h-q_l)} > 0$, we have H_1 is negatively definite. Thus, the second-order condition holds.

Therefore, the optimal price of the firm providing the different strategy simultaneously ought to meet the following first-order condition:

$$\frac{\partial \prod^{PNE}}{\partial p_l^{PNE}} = 0,\ \frac{\partial \prod^{PNE}}{\partial p_h^{PNE}} = 0 \tag{13}$$

The optimal price of the firm is represented by

$$\begin{aligned} p_h^{PNE^*} &= \frac{q_h(\gamma N^e - \lambda q_h + 1)}{2} \\ p_l^{PNE^*} &= \frac{q_l(\gamma N^e - \lambda q_l + 1)}{2} \end{aligned} \tag{14}$$

From Eq. (14) and $p_h^{PNE^*} - p_l^{PNE^*} \geq 0$, we can obtain $\gamma \geq \frac{\lambda(q_h+q_l)-1}{N^{PNE}}$.

We now proceed by requiring that the consumers' expectations are rational (fulfilled in optimal), while they cannot be influenced by the prices set by the firms. We impose rationality by substituting $N^e = N^{PNE}$ into Eq. (11) then inserting the result into Eq. (14).

Then, we can obtain the optimal total demand and prices of the firm providing two products of differentiate quality:

$$\begin{aligned} p_h^{PNE^*} &= \frac{q_h\big[2(1-\lambda q_h) + \lambda\gamma(q_h - q_l)\big]}{2(2-\gamma)} \\ p_l^{PNE^*} &= \frac{q_l(1-\lambda q_l)}{2-\gamma} \\ N^{PNE^*} &= \frac{1-\lambda q_l}{2-\gamma} \end{aligned}$$

Further, we get independent demand and profit of the firm.

$$\begin{aligned} N_l^{PNE^*} &= \frac{\lambda q_h}{2} \\ N_h^{PNE^*} &= -\frac{2\lambda(q_l+q_h) - \lambda\gamma q_h - 2}{2(2-\gamma)} \\ \pi^{PNE^*} &= \frac{q_h\big(4\lambda^2 q_l^2(\gamma-1) - \lambda^2 q_l q_h\big(\gamma^2-4\big) + \lambda^2 q_h^2(\gamma-2)^2 + 4\lambda\gamma(q_h-q_l) - 8\lambda q_h + 4\big)}{4(2-\gamma)^2} \end{aligned}$$

From $p_l^{PNE^*} > 0$, $p_h^{PNE^*} > 0, 0 < \gamma < 1, 0 < N^{PNE^*} < 1$, we can obtain $0 < \lambda q_l < \lambda q_h < 1, 0 < \gamma < 1$ or $0 < \lambda q_l < 1 < \lambda q_h < 2$, $\frac{2(\lambda q_h-1)}{\lambda(q_h-q_l)} < \gamma < 1$. ■

Proof of Corollary 1

$$\frac{\partial \pi^{WE^*}}{\partial \gamma} = \frac{(\lambda - c + c\lambda)^3}{27\lambda^4(1-\gamma)^2} > 0$$

$$\frac{\partial p_h^{WE^*}}{\partial \gamma} = \frac{2c(\lambda - c + c\gamma)}{9\lambda^3} > 0$$
$$\frac{\partial q_h^{WE^*}}{\partial \gamma} = \frac{c}{3\lambda^2} > 0$$
$$\frac{\partial N_h^{WE^*}}{\partial \gamma} = \frac{1}{3\lambda(1-\gamma)^2} > 0$$

■

Proof of Proposition 2 By setting $N^{PE} = N^e$ into Eq. (11), we can obtain

$$\begin{cases} N_l^{PE} = \frac{\lambda q_l q_h (q_h - q_l) + p_h^{PE} q_l - p_l^{PE} q_h}{q_l (q_h - q_l)} \\ N_h^{PE} = \frac{p_l^{PE} q_l - p_h^{PE} q_l + q_l q_h + \gamma q_l^3 - q_l^2 - \lambda q_l q_h^2 - \gamma p_l^{PE} q_h + \gamma p_h^{PE} q_l + \lambda \gamma q_l q_h (q_h - q_l)}{q_l (q_h - q_l)(1-\gamma)} \end{cases} \tag{15}$$

Inserting Eq. (15) into Eq. (4), we get the profit of the firm providing two types of products of differentiated quality

$$\begin{aligned} \max_{p_l, p_h} \prod^{PE} &= p_l^{PE} N_l^{PE} + p_h^{PE} N_2^{PE} \\ &= p_l^{PE} \left(\frac{\lambda q_l q_h (q_h - q_l) + p_h^{PE} q_l - p_l^{PE} q_h}{q_l (q_h - q_l)} \right) \\ &\quad + p_h^{PE} \left(\frac{p_l^{PE} q_l - p_h^{PE} q_l + q_l q_h + \gamma q_l^3 - q_l^2 - \lambda q_l q_h^2 - \gamma p_l^{PE} q_h + \gamma p_h^{PE} q_l + \lambda \gamma q_l q_h (q_h - q_l)}{q_l (q_h - q_l)(1-\gamma)} \right) \end{aligned} \tag{16}$$

Here, the Hessian matrix is as follows

$$H_2 = \begin{bmatrix} -\frac{2q_h}{q_l(q_h - q_l)}, & \frac{\gamma q_l - 2q_l + \gamma q_h}{q_l (q_h - q_l)(\gamma - 1)} \\ \frac{\gamma q_l - 2q_l + \gamma q_h}{q_l (q_h - q_l)(\gamma - 1)}, & -\frac{2q_l}{q_l (q_h - q_l)} \end{bmatrix}$$

From $-\frac{2q_h}{q_l(q_h - q_l)} < 0$, $\frac{4q_l(1-\gamma) + \gamma^2 (q_l - q_h)}{q_l^2 (q_h - q_l)(1-\gamma)^2} > 0$, we have H_2 is negatively definite. Thus, we have $4q_l(1-\gamma) + \gamma^2(q_l - q_h) > 0$.

Therefore, the optimal price of the firm providing the differentiated strategy simultaneously ought to meet the following first-order condition:

$$\frac{\partial \prod^{PE}}{\partial p_l^{PE}} = 0, \frac{\partial \prod^{PE}}{\partial p_h^{PE}} = 0 \tag{17}$$

The optimal prices of the firm is represented by

$$p_h^{PE^*} = \frac{q_l q_h (1-\gamma)\left[2(1 - \lambda q_h) + \lambda \gamma (q_h - q_l)\right]}{4q_l(1-\gamma) + \gamma^2 (q_l - q_h)}$$

$$p_l^{PE^*} = \frac{q_l\big(\lambda\gamma\big(q_l^2+q_h^2\big)+\lambda\gamma^2 q_h(q_l-q_h)-2\lambda q_l^2-\gamma(q_l+q_h)+2q_l\big)}{4q_l(1-\gamma)+\gamma^2(q_l-q_h)} \tag{18}$$

Inserting Eq. (18) into Eqs. (15) and (16), we have the optimal demand and profit are

$$\begin{aligned}
N_l^{PE^*} &= \frac{q_h(\gamma+2\lambda q_l(1-\gamma)-\lambda\gamma q_h)}{4q_l(1-\gamma)+\gamma^2(q_l-q_h)}\\
N_h^{PE^*} &= \frac{-(\gamma-2+2\lambda(q_l+q_h)-\lambda\gamma q_l-2\lambda\gamma q_h)q_l}{4q_l(1-\gamma)+\gamma^2(q_l-q_h)}\\
N^{PE^*} &= \frac{2q_l(1-\lambda q_l)+\gamma(q_h-q_l)[1-\lambda(q_l+q_h)]}{4q_l(1-\gamma)+\gamma^2(q_l-q_h)}\\
\pi^{PE^*} &= \frac{q_l q_h\big(1+\lambda^2\big(q_h^2-q_l^2+q_l q_h\big)-2\lambda q_h+\lambda^2\gamma\big(q_l^2-q_h^2\big)-\lambda(q_l-q_h)\big)}{4q_l(1-\gamma)+\gamma^2(q_l-q_h)}
\end{aligned}$$

From $0 < N^{PE^*} < 1$, $p_l^{PE^*} > 0$, $p_h^{PE^*} > 0$, $4q_l(1-\gamma)+\gamma^2(q_l-q_h) > 0$, we can obtain

$$0 < \lambda < \frac{3\gamma q_l-2q_l+\gamma q_h+\gamma^2 q_h-\gamma^2 q_l}{\gamma\big(q_h^2-q_l^2\big)+2q_l^2}, 4q_l(1-\gamma)+\gamma^2(q_l-q_h) > 0$$

$$\begin{cases} 1-\lambda q_h < 0, \dfrac{2q_l(\lambda q_h-1)}{\lambda(q_h-q_l)} < \gamma < 1 \text{ or} \\ 1-\lambda q_h > 0, 1-\lambda(q_l+q_h) < 0, 0 < \gamma < \dfrac{2q_l(1-\lambda q_l)}{(q_h-q_l)[\lambda(q_l+q_h)-1]} \end{cases}$$

■

Proof of Proposition 3 By solving the first-order condition $\frac{\partial U}{\partial q_h} = 0$, we can obtain

$$q_h^*(\theta) = \frac{\theta+\gamma N^e}{2\lambda}$$

According to $\frac{\partial q_h(\theta)}{\partial \theta} > 0$, we can get

$$q_h^*(\theta) = q_h^*(\theta_h) = \frac{\theta_h+\gamma N^e}{2\lambda} \tag{19}$$

Combining Eqs. (6) and (19), we can obtain

$$p_h^* = \frac{(\theta_h+\gamma N^e)^2}{4\lambda} = \lambda q_h^2 \tag{20}$$

Insetting Eq. (10) $\theta_h = \frac{p_h-p_l}{q_h-q_l} + \lambda(q_h+q_l) - \gamma N^e$ into Eq. (19), we have

$$p_h - p_l = \lambda(q_h - q_l)^2 \tag{21}$$

According to Eqs. (20) and (21), we can obtain $p_h = \lambda q_h^2$, $p_l = \lambda q_l(2q_h - q_l)$. ■

Proof of Corollary 2

$$p_h^{WNE^*} - p_h^{WE^*} = \frac{(\lambda - c + c\gamma)^2}{\lambda^3(3-\gamma)^2} - \frac{(\lambda - c + c\gamma)^2}{9\lambda^3} > 0$$

$$q_h^{WNE^*} - q_h^{WE^*} = \frac{\lambda - c + c\gamma}{\lambda^2(3-\gamma)} - \frac{\lambda - c + c\gamma}{3\lambda^2} > 0$$

$$N_h^{WNE^*} - N_h^{WE^*} = \frac{\lambda + 2c}{\lambda(3-\gamma)} - \frac{\lambda + 2c - 2c\gamma}{3\lambda(1-\gamma)} < 0$$

$$\begin{aligned}\pi_h^{WNE^*} - \pi_h^{WE^*} &= \frac{(\lambda - c + c\gamma)^3}{\lambda^4(3-\gamma)^3} - \frac{(\lambda - c + c\gamma)^3}{27\lambda^4(1-\gamma)} \\ &= \frac{\gamma(\gamma-6)(\gamma+3)(\lambda - c + c\gamma)^3}{27\lambda^4(3-\gamma)^3(1-\gamma)} < 0\end{aligned}$$

■

Proof of Proposition 4 is similar to that of Proposition 1. Proof of Proposition 5 is similar to that of Proposition 2. Thus, we omit it for brevity.

References

1. B. Jing, Network externalities and market segmentation in a monopolist. Econ. Lett. **95**(1), 7–13 (2007)
2. M. Li, H. Feng, F. Chen, Optimal versioning and pricing of information products with considering or not common valuation of customers. Comput. Ind. Eng. **63**(1), 173–183 (2012)
3. M. Li, H. Feng, F. Chen, J. Kou, Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers. Comput. Oper. Res. **40**(10), 2374–2386 (2013)
4. Z. Liu, M. Li, J. Kou, Selling information products: sale channel selection and versioning strategy with network externality. Int. J. Prod. Econ. **166**, 1–10 (2015)
5. H.K. Cheng, Q.C. Tang, Free trial or no free trial: optimal software product design with network effects. Eur. J. Oper. Res. **205**(2), 437–447 (2010)
6. S. Shivendu, Z. (James) Zhang, Versioning in the software industry: heterogeneous disutility from underprovisioning of functionality. Inf. Syst. Res. **26**(4), 731–753 (2015)
7. R.K. Chellappa, S. Shivendu, Mechanism design for ‘free’ but ‘no free disposal’ services: the economics of personalization under privacy concerns. Manage. Sci. **56**(10), 1766–1780 (2010)
8. R.K. Chellappa, A. Mehra, Cost drivers of versioning: pricing and product line strategies for information goods. Manage. Sci. **64**(5), 2164–2180 (2018)
9. B. Jing, Lowering customer evaluation costs, product differentiation, and price competition. Manage. Sci. **35**(1), 113–127 (2016)

10. K. Griva, N. Vettas, Price competition in a differentiated products duopoly under network effects. Inf. Econ. Policy **23**(1), 85–97 (2011)

The Comparison of Two Styles of Residence Layout Examined by Sunshine—Beijing and Barcelona

Yunan Zhang, Chun Ji, Shuxuan Xing, and Hui Zhou

Abstract Good sunshine conditions are free disinfection and cleaning methods for urban buildings and outdoor spaces. The application of today's simulation software and related sunshine analysis programs can already analyze sunshine conditions very accurately. This paper attempts to calculate according to the same sunshine standard, and compares the sunshine conditions between the building surface and public space of China's currently relatively common north–south high-rise buildings and multi-storey buildings represented by Beijing, and the oblique, small street area and dense road network urban space represented by Barcelona, and the conditions between buildings and urban space, so as to obtain a regular understanding of how to strengthen the uniformity of sunshine conditions in the process of building and urban group layout, thus providing reference suggestions for strengthening the building and urban layout.

Keywords Cerda style of Barcelona · Beijing style of residence · Sunshine examination · Sunshine standard

1 Two Different Styles of Layouts and Their Influence

Beijing and Barcelona are both international metropolises, with one in the east of Eurasia land and the other in west, but now they have different city layouts. Beijing adopts a north–south street network, with wide roads and large blocks. Most of the congregated dwelling under construction now adopt the form of single-high-rise

Y. Zhang · C. Ji (✉) · S. Xing · H. Zhou
iCIR of Beijing Jiaotong University, Beijing Jiaotong University, Beijing, China
e-mail: 2911847511@qq.com

Y. Zhang
e-mail: 910460222@qq.com

S. Xing
e-mail: 594036859@qq.com

H. Zhou
e-mail: 1339829554@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_48

buildings under 80 m. Barcelona, on the other hand, uses small streets and dense road networks. The streets grid twists an angle of 45° to the north–south direction. Most of them use 7–9 layers of multi-storey buildings to surround the streets. The urban fabrication of the two cities have respectively become the representatives of the two different planning patterns in the east and west.

Due to the requirement to get sunshine, Beijing's residential space lacks east–west enclosure, and the buildings between the north and south need to maintain large distances, which also limits the floor area ratio of residential buildings. While, the pleasant scale of Barcelona's urban space attracts the world's attention and is widely praised by urban researchers from all over the world. Under Cerda's planning pattern, Barcelona's narrow streets are very humane, and the small blocks make the road network dense and traffic smooth.

Many Chinese researchers had imaged that the difference in geographical environment and climate between Beijing and Barcelona is the root of the formation of two different layouts. Other scholars believe that the different sunshine standards of dwelling buildings in China and Spain make the streets of Barcelona consider lack of the influence of sunshine. Even some scholars believe that the 45° twist of Barcelona's street grid is caused by factors such as the restriction of coastline. Therefore, this study hopes to calculate according to the same sunshine standard and discuss the advantages and disadvantages of the two kinds of layouts.

Barcelona's road network is about 133.3 m^2, and the typical street width is 20 m, that is to say, the distance between the walls on both sides of the road is 20 m, and the middle is about 10 m for motor vehicle lanes and roadside parking lots. The buildings in the downtown named Eixample districtare usually 7 floors, but in some cases it is possible to add 2 floors and the top floor reached to 9 floors. The building has a chamfered design at the corner to form an octagonal corner open space. Some data show that the depth of residential areas in Barcelona is larger, usually about 24 m, and some can reach more than 30 m with the help of courtyard in the middle of buildings, such as the famous casa MILà. All most all of European buildings are full of restaurant and retail on the ground floor.

The plot of land in Beijing is not usually so square, sometimes even in polygonal shape, but the density of most street networks is not more than 350*350 m. At present, due to the government's regulations of house, floor area ratio and height control become two limit factors, it is about 80 m (26 floors) below and the floor area ratio does not exceed 3.1. There are also regulations of -high-rise buildings distance that need to have no less than 13-m to prevent the spread of fire, and the current building must not be wider than 80 m, on this conditions, increasing the depth of building is an efficient way to increase the plot ratio of land. As patios are not allowed to use, the depth of Beijing slab houses is usually about 15 m. In general, except for few public service room, Beijing's residential areas are all residential function from top to bottom buildings (Reference: Fengtai residential building design guidelines).

2 The Sunshine Influence Factors and Its Deepening Analysis Method

The sunshine standard commonly used in Beijing is no less than two hours for main room on Great cold day which is Jan 20 for every year, and the effective calculation period is from 8:00-am to 16:00-pm. Many sunshine design software can simulate the sunshine impact of buildings. The principle is that after inputting the latitude and effective sunshine duration of the building, the computer is responsible for point-by-point calculation of the area affected by different incident angles of sunlight at different times to judge the sunshine duration at each grid point. This method of calculating densely dotted points based on mathematical geometric factors has a very accurate result and has been widely used in the field of architectural design.

Latitude is one of the key factors in calculating sunshine. Contrary to what many Chinese scholars had thought, although Barcelona is relatively warm under the influence of Mediterranean climate, the latitude of Barcelona is further north than Beijing, almost close to Shenyang city in China. However, the warm climate itself has nothing to do with the calculation results of building sunshine, so we need to consider the latitude of Beijing or Barcelona to achieve the comparison of the two methods. In this paper, the latitude of Beijing is used as the basis for simulation, and the actual verification results will be more relaxed than Barcelona's conditions in fact. It can also be adapted to the area in most parts of northern China.

Based on our understanding of the layout of two different cities, we should not only calculate the floor area ratio according to the actual 30 m commonly used in many residential buildings in Barcelona and Cerda's initial ideal depth of 24 m, but also consider the calculation of floor area ratio for residential buildings with a depth of about 15 m under the Beijing model without patios. However, Beijing's residential buildings are in the form of high-rise buildings from the north to the south. The depth of 15 m has reached the largest scale and cannot be expanded any more.

At the same time, considering Cerda Barcelona's block building layout, all buildings can be frontage, so some corners with less sunshine effect can increase office space moderately to form a balanced community (Fig. 1). However, in Beijing's plan, although residential buildings can be transformed into office use in the same way, they are rarely occur in practice. Therefore, this part of Barcelona needs special marking.

Another research result of the project team shows that people believe that the city streets and buildings also need better sunshine conditions, and the city streets not only need better sunshine in winter, but also need better shading effect in summer, so as to be more conducive to providing shade for people and reducing the heat island effect of big city. So the sunshine conditions on the hot summer days also should be calculated, but at this stage, we need to consider the shade area of street trees and the shadow of the house itself to prevent from the sun. The spacing of street trees is 8 m and the crown diameter is 8 m, these tree are array along the street and even at the chamferd corner.

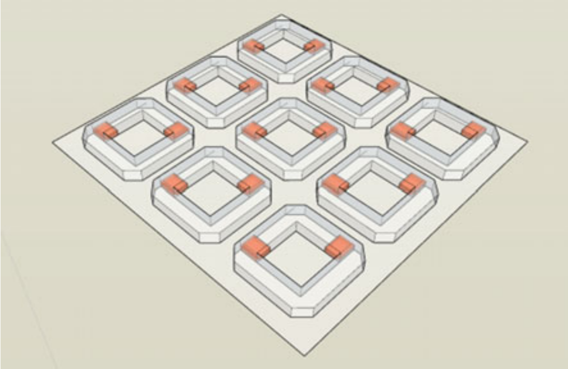


Fig. 1 Office-residence balance

3 The Results of Two Different City Layout

Through computer simulation and calculation, the sunshine level at the east and west corners of the block is the worst when the Barcelona style is 15, 24 and 30 m as shown in Fig. 2. Due to the 45 included angle, the street width of 20 m can basically meet with the sunshine needs of 6-storey buildings ensure the sunshine level of street. If the top floor is retreated, 9-storey buildings can also match the street of this scale.

By examining the layout of Beijing, due to the north–south building direction, the back row buildings need to step back a very long distance. Although high-rise buildings are adopted, the floor area ratio of the buildings is basically controlled to about 3.1 on the whole district. However, there are more areas in the shadow of building lack of sunshine in the east–west streets (Fig. 3).

The results of variables such as floor area ratio and street width under different depths and layouts are as follows (Table 1):

Through comparison, it is found that the building density of Barcelona layout is not lower than that of Beijing style even under the same sunshine conditions in winter, and if Barcelona's existing patio layout is adopted, it will greatly exceed the current floor area ratio of Beijing layout. In the Cerda's layout, even though the

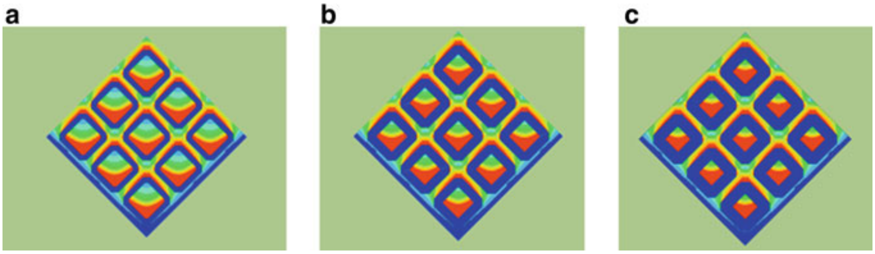


Fig. 2 **a** 15 m, **b** 24 m, **c** 30 m

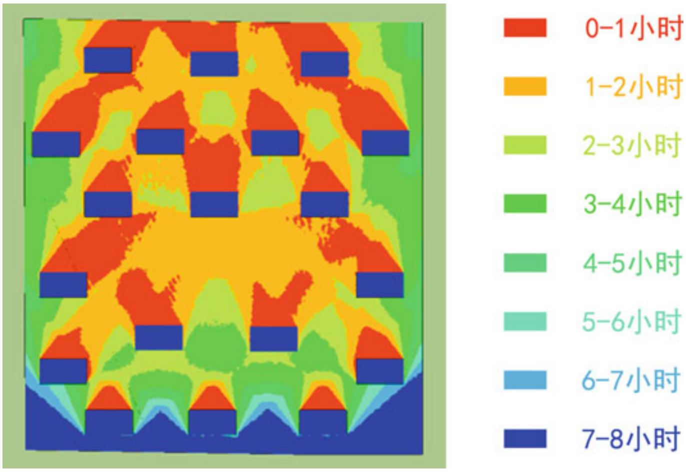


Fig. 3 The layout of Beijing

Table 1 FAR and street width under different depths and layouts

Road width	Depth		
	15 m	24 m	30 m
17 m	FAR:3.02	FAR:4.49	FAR:5.21
20 m	FAR:3.08	FAR:4.52	FAR:5.29
24 m	FAR:3.17	FAR:4.62	FAR:5.40

streets are narrow, they are generally able to receive sufficient sunshine in winter. In the determinant layout of Beijing, the two rows of buildings need a larger distance between the north and the south, which increases the width of the streets accordingly and makes the east–west streets have poor sunshine conditions in winter.

Looking at the sunshine in the atrium space in the Barcelona layout, only the bottom within 20 m of the corner does not meet the sunshine requirement (Fig. 4), but because it is close to the corner, it can be used for office and other commercial purposes (Fig. 5).

By comparing the sunshine conditions of the primary and secondary sunlight receiving surfaces of the buildings themselves, it is found that the rooms towards

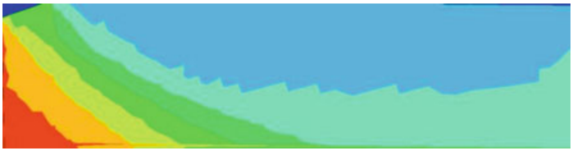


Fig. 4 Sunshine in the atrium space

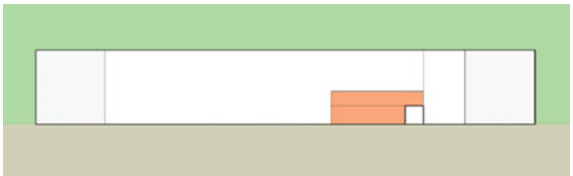


Fig. 5 For office and other commercial purposes

Fig. 6 The rooms towards north in Beijing building

north in Beijing building have no sunshine in winter (Fig. 6), while the buildings with secondary light receiving surfaces towards north east or north west directions arranged in Barcelona can also get more than one hour of sunshine in winter. For many rooms located on the secondary light receiving facade, the quality improvement effect is obvious.

At the same time, according to the simulation of sunshine angle, it is found that because of the southeast and southwest sunshine, the sunshine in winter enters the indoor far away from the outer wall, while the sunshine incidence angle of Beijing layout is higher and the indoor light receiving range is much smaller (Fig. 7).

Comparing the sunshade effect of buildings in summer, it is found that the overall sunshine time of the building surface in Beijing is longer during the summer heat, while the sunscreen effect in Basesselda's layout is not good. However, with the use of Blinds and some other advanced sunshade facilities, the problem of sunshade inside the building can be fully solved (Fig. 8).

After comparing the sun protection and shading effects of streets in summer, and comparing the shading conditions of two layouts in summer, it is found that street trees and building combinations can be covered in the Barcelona-Cerda's layout 60%

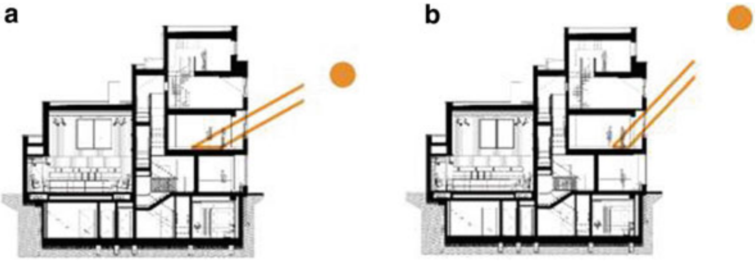


Fig. 7 **a** Barcelona layout. **b** Beijing layout

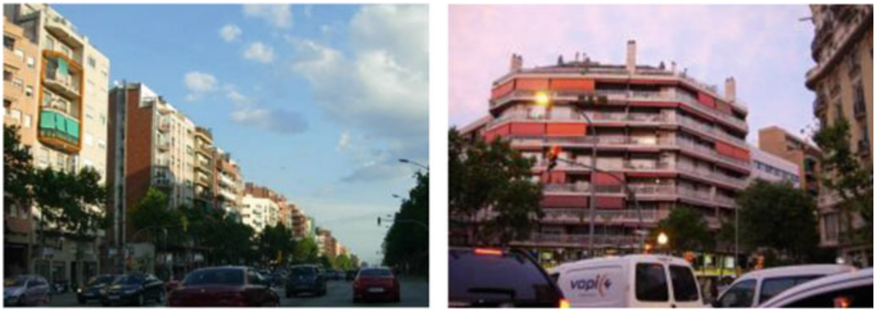


Fig. 8 Sunshade facilities

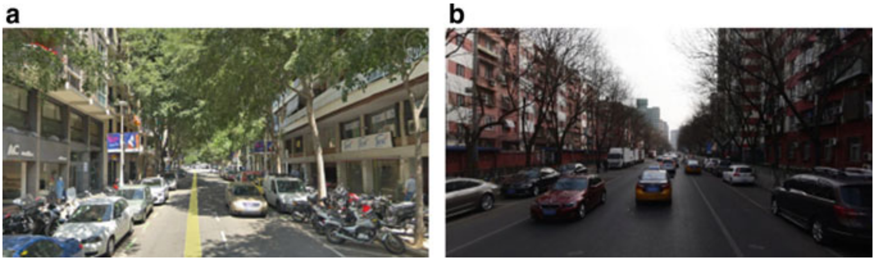


Fig. 9 Barcelona. **b** Beijing

of the road area, and in Beijing layout due to the wide roads, green coverage rate is much smaller, which is only 33% (Fig. 9).

By comparing and analyzing the traffic of the two layouts, it is found that the planned road network in Barcelona-Zelda is much dense and convenient for traffic microcirculation, while the Beijing model requires large blocks, has low road network density and is prone to concentrated congestion points.

In addition, considering the layout of buildings along the street, Barcelona has more space along the street which is more easily to make commercial spaces and enhance the value of ground floor, which is convenient for the mixed use of the community and the balance of occupation and residence, and is also conducive to the logistics transportation of buildings along the street. But need to improved social management. In addition, the individual space in the corner part can be vertically separated from other parts for office or commercial purposes, and its location on blocks and value is also very well.

4 Conclusions and Prospect

The comparison between Beijing determinant layout and Barcelona Zelda layout can draw the following conclusions: Despite factors such as warm climate, there is not much difference in floor area ratio between Barcelona's multi-storey layout and Beijing's high-rise layout, and even the floor area ratio can be higher. The 45° twist with the positive north–south direction is the key factor to form a small street area and dense road network. From the perspective of sunshine in winter and sun shading in summer, Barcelona's overall effect is better than Beijing's layout, which sticks to the north–south orientation of individual houses. It also brings many benefits to the development of mixed communities, but more efforts are needed in social management. In the future planning and design, some residential districts in northern China can try to use Barcelona layout for reference, which can not only enrich the space form but also enhance the sunshine intensity, and is of great significance to the improvement of sanitary conditions and comfort level of residential districts. In addition, it can increase the degree of functional mix in the region and form a job-residence balance.

References

1. J. Hang, The thinking and system of Cerda's Barcelona Eixample plan [D]. Zhejiang University (2012)
2. J. Busquets, A. Lu, Y. Xu, Urban history as clue for designing the cities today: the case of Barcelona and its grid plan of Cerda [J]. Architectural J. (11), 2–16 (2012)
3. J. Busquets, L. Qian, Barcelona: plans for the metropolis. Following a systemic approach [J]. Architect (1), 31–41 (2018)
4. F. Hang, The renovation of Barcelona city [J]. Architectural J. **5**, 57–61 (2002)
5. C. Li, The study on the culture and spacial development strategy of Barcelona [J]. Urban Plann. Overseas **4**, 67–71 (2004)
6. L. Dai, S. Gai, Study on the new town construction and old city regeneration of Barcelona in the turning point of mega-events effect [J]. Chin. Overseas Architecture (02), 59–62 (2009)
7. J. Zhang, L. Yuan, The enlightenment of the development of multi-rise high density perimeter blocks in Barcelona's Example [J]. Urban Plann. Overseas **04**, 51–55 (2004)
8. J. Busquets, *Barcelona: Urban Evolution of a Compact City* [M] (Actar, Barcelona, 2006)

Manpower Planning for MTR Carriage Assembly with an Integer Programming

Yingjiang Wu and Ray Y. Zhong

Abstract This research is motivated by a real-life case which runs the public transportation services in Hong Kong, China. The mass transit railway (MTR) is contemplating the maintenance by considering the manpower planning in its depot. The main purpose of this research is to improve the efficiency of assembly with adjustable manpower planning based on an integer programming optimization model based on specify assumptions is presented. The strength of this model simplifies practical problems by considering several scenarios in the MTR assembly operations. The numerical study shows that the model of this paper can be effectively used to improve the utilization and quality of the practical applications.

Keywords Manpower planning · Carriage assembly · Integer programming

1 Introduction

The Mass Transit Railway (MTR) in Hong Kong as a significant public transportation service provider, supports over eight million of residents to have a convenient daily-life [1]. Thereby, a complete, and progressive carriage maintenance system should be a footstone for providing such large volume of transportation. Such maintenance is highly labor-intensive. The arrangement of fixed manpower in labor-intensive industries can not only optimize the allocation of resource, but also keep down for corporations [2]. Manpower assignment should be designed for different stages, which follow the priority: (1) disassembly, inspection, and assembly. Efficiency in assembly step will be principally considered in this paper.

Firstly, assembly method is a manufacturing process, in which items are created in a sequential manner [3]. Team production method as a development assembly

Y. Wu (✉)
Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
e-mail: 17250315@life.hkbu.edu.hk

R. Y. Zhong
Department of Industrial and Manufacturing Systems Engineering, University of Hong Kong, Hong Kong, China
e-mail: zhongzry@hku.hk

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_49

line approach, is exploring the equality between demand and supply in manpower, which implies lower cost of labor with insured normal maintenance. In reality, MTR has already designed a project team called ‘PjDiv’, with mixture of people, size of team for coping with dynamic problems especially for maintenance tasks [4]. Hence, turnover in distinct teams and shift cost should be considered.

Secondly, rolling stock assembly workload can be divided into two parallel steps, bodyshell assembly and bogie assembly, and gather at the fitting out. More detailly, bodyshell parts, include underframe, sides, and roof of the carriage, also corresponds to bogie part, which includes brake rigging, wiring and the wheelsets [5].

This paper uses a typical integer programming to examine the manpower planning [6]. There are three main methods currently being adopted to solve integer programming, branch and bound method, zero-one, and implicit enumeration, with the basic solving flow such as decomposition, relaxation, and exploration [7]. And Zero-One as a special case of branch and bound method.

Finally, those preconditions refer to changeable issues, like manpower hourly cost, number of unattended vehicles, and working hours should be fixed in advance to refine from complication as another significant aspect.

The mass transit operators planning has long been a question of great interest in a wide range of fields, such as Ismaila et al. [8], Chen et al. [9], Zhang et al. [10], Karthikeyan and Krishnaswamy [11], and Jaillet et al. [12]. Chen et al. found that mixed integer programming can be useful in manpower supply by comparing basic stochastic model and integrated stochastic model [9]. In order to compare the weakness and strengths between integer programming (IP) and mixed integer linear programming (MILP), Khoshniyat and Krasemann have provided sufficient evidences to prove that MILP can generate good solutions in scheduling as well [13]. However, researchers have not treated operator’s assignment in much details. A much-debated question is whether their model still works for assembly procedures, given with aspects which are mentioned in the previous paragraphs.

The purpose of this paper is to construct a model to minimum manpower cost in completing assembly tasks, given with fixed manpower and daily tasks. The remaining part of this paper proceeds as follow: Sect. 2 states the problem description, given with assumptions, parameters, variables and the details of model, including objective function, constrains and their meaning explanations. Others like solving procedures, and even results, will be shown in Sect. 3. Section 4 not only concludes this paper, but also highlights the contribution, limitation and future works.

2 Problem Description and Formulation

In this section, a MIP model will be given for curtailing manpower cost and scheduling the shift between in several tasks. Problem description would be specified. Assumptions and parameters are described in Sects. 2.2 and 2.3, respectively. Fourthly, the objection function and constrains are demonstrated from scheduling aspects.

2.1 Problem Description

Figure 1 shows that the whole maintenance procedure in repair factory, which chiefly highlights the assembly operation. In whole maintenance stage, an unfixed rolling stock should completely experience the previous procedures, disassembly and inspection, before stepping into assembly operations. Based on the purpose of this paper, research principally focuses on the assembly operation of carriage.

In assembly operation, there are two parallel sub-parts of assembly require to be done, which are bogie and bodyshell of carriage, respectively. Each of sub-operations is respectively represented by I and J, to classify these sub-operations in assembly. To superior understand task stage, each sub-operation also can be divided into three distinct categories for next stage.

In order to quickly and efficiently tasks under each sub-operation, Ii and Jj, are symbols for each task, given with exact i and j. In this system, the tasks under body-shell sub-operation can be shown as {I1 = underframe, I2 = sides, I3 = roof of the carriage}. Corresponding to tasks of body-shell, the tasks of bogie operation can be best treated under three terms as well, given with {J1 = brake rigging, J2 = wiring, J3 = wheelsets}.

Corresponding with each task, manpower allocation and shift happen during repairing period. What senior level considers at this stage is minimum the manpower cost and shift cost for their corporates. Meaningfully, utilization is maximized when the cost is minimized.

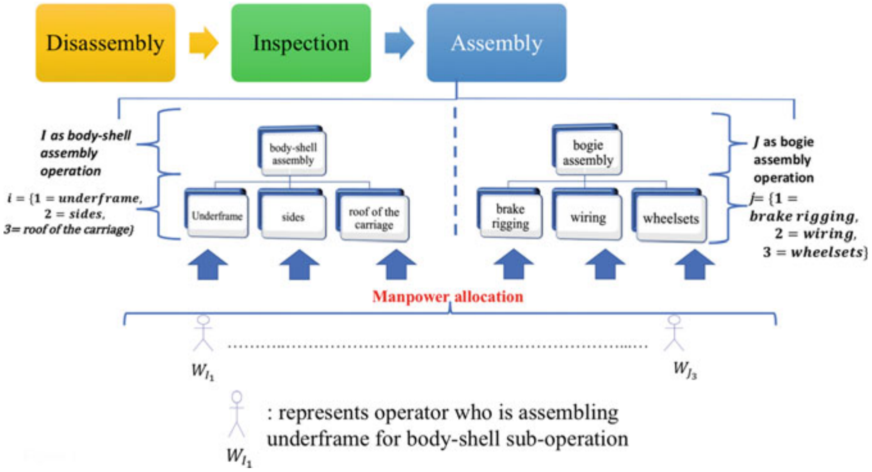

Disassembly
Inspection
Assembly
I as body-shell assembly operation
i = {1 = underframe, 2 = sides, 3= roof of the carriage}
body-shell assembly
Underframe
sides
roof of the carriage
bogie assembly
brake rigging
wiring
wheelsets
J as bogie assembly operation
j= {1 = brake rigging, 2 = wiring, 3 = wheelsets}
Manpower allocation
W_{I_1}
W_{J_3}
: represents operator who is assembling underframe for body-shell sub-operation
W_{I_1}

Fig. 1 Problem specification

2.2 Assumptions

In order to solve this problem, several assumptions are made as follows:

1. Only one carriage is operated during the whole procedure.
2. Daily working hours of each operator should be less or equal to 8 h. And assembly ability of operators can keep stable before next shift.
3. Each shift stably keeps 4 h in each time. The gap of twice shifts for one person should be at least 4 h and at most twice shifts one day.
4. No priority list of operators for dispatching.
5. Operators only focus on one specified task before next shift time, which implies that they cannot interchange during working time.
6. Manpower salary is based on what they do.
7. Each task can process parallel and continuously.
8. Quality of previous double operation, disassembly and inspection, should be guaranteed. But there will be different scenarios based on random procedure situations.
9. Costs of other issues are ignorable. Expense only comes from shift cost and normal working salary and are hourly counted.
10. Employees only need to work a five-day workweek with two consecutive days off.

2.3 Symbols

2.3.1 Parameters

I	The set of all tasks in body-shell assembly
J	The set of all tasks in bogie assembly
st	The setup time of task
p_i	The setup operator's assignment of task for body-shell assembly
p_j	The setup operator's assignment of task for bogie assembly
D	Week day set, where {0: Monday, 1: Tuesday, 2: Wednesday, 3: Thursday, 4: Friday}
d	Element of D set represent the working day and $d \in$ D
T	Daily time set hourly corresponds to 24 h in each day. For instance, 0 of set T represents 0 o' clock
t	Element of T set, hourly time in one day
ST_1	Starting time of the task in first shift
ST_2	Starting time of the task in second shift
θ	Random scenarios of the task identify insufficient manpower
P	The set of type P teams
m_p	Number of persons in type P teams

pw_1 Regular manpower cost (hourly per person) when supply satisfies demand for body-shell assembly

pw'_1 Regular manpower cost (hourly per person) when supply satisfies demand for bogie assembly

pw_2 Excess manpower cost per person for body-shell assembly

pw'_2 Excess manpower cost per person for bogie assembly

PW_3 The set of manpower cost in different time for body-shell assembly

PW'_3 The set of manpower cost in different time for bogie assembly

pw_3 Insufficient manpower cost per person at t time of d date for body-shell assembly, and $pw_3 \in PW_3$

pw'_3 Insufficient manpower cost per person at t time of d date for bogie assembly, and $pw'_3 \in PW'_3$

$E(\theta)$ The expected value of scenarios θ.

2.3.2 Variables

c Hourly shift cost of the task per person.

w_{dt} Total shift time operators at t time of d date for the task.

s_{dt} Total used shift hours for tasks, which starts at the t time of d date.

h Total needed hours in task.

v_{dtpj} Number of types P team starting from t time in d date for bogie assembly.

v_{dtpi} Number of types P team starting from t time in d date for body-shell assembly.

a_{tdi} Number of excess manpower in t time of d date for the task of body-shell assembly.

a_{tdj} Number of excess manpower in t time of d date for the task of bogie assembly.

$\alpha_{tdi}(\theta)$ $pw_3 \times \mathrm{E}\left(\sum_{i=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdi}(\theta)\right)$ represents the expected value for the insufficient manpower for the body-shell assembly for all scenarios in t time of d date.

$\alpha_{tdj}(\theta)$ $pw'_3 \times \mathrm{E}\left(\sum_{j=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdj}(\theta)\right)$ represents the expected value for the insufficient manpower for the bogie assembly for all scenarios in t time of d date.

x_{dti} x = 1 if there are shift starting from t time of d date for the task of body-shell assembly, and 0 otherwise.

x_{dtj} x = 1 if there are shift starting from t time of d date for the task of bogie assembly, and 0 otherwise.

B A large value for case of modelling.

2.4 Formulation and Its Explanation

The following formulation is constructed based on the idea of Chen et al. [9] by considering shift cost and regular manpower cost.

$$\text{Min}\, z = pw_1 \sum_{i=1}^{3} \sum_{p \in P} \sum_{d \in D} \sum_{t \in T} hm_p v_{dtpi} + pw_1' \times \sum_{j=1}^{3} \sum_{p \in P} \sum_{d \in D} \sum_{t \in T} h \times m_p \times v_{dtpj} + pw_2 \times$$
$$\sum_{i=1}^{3} \sum_{d \in D} \sum_{t \in T} a_{dti} + pw_2' \times \sum_{j=1}^{3} \sum_{d \in D} \sum_{t \in T} a_{dtj} + pw_3 \times \mathrm{E}\left(\sum_{i=1}^{3} \sum_{d \in D} \sum_{t \in T} \alpha_{tdi}(\theta) \right) + pw_3'$$
$$\times\, \mathrm{E}\left(\sum_{j=1}^{3} \sum_{d \in D} \sum_{t \in T} \alpha_{tdj}(\theta) \right) + \left(c \times (\sum_{d \in D} \sum_{t \in T} w_{dt} s_{dt}) \right)$$

Subject to,

$$\sum_{i=1}^{3} st_i + \sum_{j=1}^{3} st_j \le (\sum_{i=1}^{3} h_i + \sum_{j=1}^{3} h_j) + s_{dt} \tag{1}$$

$$\left(\sum_{i=1}^{3} m_p \times v_{dtpi} \right) - a_{tdi} + \mathrm{E}\left(\sum_{i=1}^{3} \sum_{d \in D} \sum_{t \in T} \alpha_{tdi}(\theta) \right) = \sum_{i=1}^{3} p_i \tag{2}$$

$$\left(\sum_{j=1}^{3} m_p \times v_{dtpj} \right) - a_{tdj} + \mathrm{E}\left(\sum_{j=1}^{3} \sum_{d \in D} \sum_{t \in T} \alpha_{tdj}(\theta) \right) = \sum_{j=1}^{3} p_j \tag{3}$$

$$ST_2 - ST_1 \ge 4 \tag{4}$$

$$s_{dt} \le 8 \times w_{dt} \tag{5}$$

$$\sum_{i=1}^{3} \sum_{d \in D} \sum_{t \in T} \sum_{p \in P} v_{dtpi} \le Bx_{dti}, \quad d \in D \,\text{and}\, t \in T \tag{6}$$

$$\sum_{j=1}^{3} \sum_{d \in D} \sum_{t \in T} \sum_{p \in P} v_{dtpj} \le Bx_{dtj}, \quad d \in D \,\text{and}\, t \in T \tag{7}$$

$$\sum_{i=1}^{3} \sum_{t \in T} \sum_{d \in D} a_{tdi} \ge 0 \in \mathbb{Z} \tag{8}$$

$$\sum_{j=1}^{3} \sum_{t \in T} \sum_{d \in D} a_{tdj} \ge 0 \in \mathbb{Z} \tag{9}$$

Table 1 Sets

Sets	Coding
T	Clock
D	Days

$$\mathrm{E}\left(\sum_{i=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdi}(\theta)\right)\geq 0\in\mathbb{Z} \tag{10}$$

$$\mathrm{E}\left(\sum_{j=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdj}(\theta)\right)\geq 0\in\mathbb{Z} \tag{11}$$

$$m_p, w_{dt}\geq 0\in\mathbb{Z}, h, s_{dt}\geq 0\in\mathbb{Z} \tag{12}$$

The objective function considers the total system cost, including the regular manpower cost, the negative surplus value of excess manpower for the specified task, the temporary manpower supply cost for insufficient power, and shift cost for the whole assembly procedure. Equation (1) of constraints states all tasks should be finished in the scheduled periods. Equations (2) and (3), as a constrain conditions of operator's assignment, indicate that original scheduled manpower, subtracting surplus workers and adding supplied operators, which should be satisfied the demand of manpower. In order to maintain third assumption, the duration of twice shifts and total used time of shift, are respectively controlled by Eqs. (4) and (5). And continued Eqs. (6) and (7) ensure operator's assignment of two sub-operations is operated only if shift exists. Equations (8), (9), (10), (11) and (12) are integer constraints for the variables.

The starting time of each shift is significant to consider in this paper. According to Chan el al., there still need a term to denote distinct the cost resulting from different starting shift time, which are the fifth and sixth terms of objective function. And it is also noted that $E\sum_{i=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdi}(\theta)$ and $E\sum_{j=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdj}(\theta)$, as expected value of number of sub-operators during t time of d date in θ scenarios, have equivalently utility with $\sum_{i=1}^{3}\sum_{t\in T}\sum_{d\in D}a_{tdi}$ and $\sum_{j=1}^{3}\sum_{t\in T}\sum_{d\in D}a_{tdj}$ in objective function (Table 1).

3 Numerical Study and Discussions

This section describes a numerical study and discussions respectively. In terms of numerical study, complete experimental environment, inputs of each parameters and variables in model are demonstrated. Additionally, a more detailed discussions on modeling results is given.

3.1 Numerical Study

Before proceeding to examine the model of this paper, the experimental environment of modeling as preconditions of programming is illustrated. The supporting experimental environment is a computer with macOS Mojave version 10.14.2, which has 1.8 GHz Intel Core i5 processor. The programming platform is Lingo with a version of 18.0.56 where the optimization software of model is conducted.

The inputs of each parameters are distributed in corresponding with practical items as close as possible through Lingo programming. Initially, regular salary for each operator preserves $35.5 per hour that is enquired by Labor Department of HKSAR for the lowest hourly reward. And hourly salary per person can also fluctuate up or down to $45 and $15 for temporary insufficient manpower and surplus manpower cases correspondingly.

In these cases, the hourly salary of insufficient manpower case and surplus manpower case is $80.5 and $50.5 respectively. And the shift hourly cost per person is $25. Total tasks, such as surplus manpower, regular manpower, and shifts, should offer same salary to their operators. Secondly, valuation of parameters is designed by developers. The detailed inputs of used time for each tasks and shifts are demonstrated in Table 2. Finally, based on the previous setting of setup time, each task needs 150 people for satisfying demands. Therefore, for instance, each operator of tasks would be paid for specified salary (i.e., $35.5 for regular situation) when the operator's assignment of each task just satisfies their demand.

In order to solve the IP, variables of model principally represent the operator's assignment in each task and shifts. Details of variables have been established Table 3. Here developers use Ei and Ej to describe the operator's assignment of surplus manpower situation as a special case. Because the Lingo system cannot support programming for calculating the expected value. Therefore, the developer desires to use Ei and Ej to define the operator's assignment of surplus manpower situation as complete terms.

3.2 Discussions

This subsection summarizes the performance of model from three vary aspects, valuation of variable and parameters, deviation of constraints and reliability of model in real, given with the outputs by Lingo statements.

3.2.1 Valuation of Parameters

(a) *Setting of hourly salary for distinct cases and shifts*

The model chose $35.5 per person hourly since companies in Hong Kong should adhere to the rule minimum hourly salary from labor department HKSAR [14].

Table 2 Parameters

Parameters		Coding		Description
Starting time	ST_1	ST1		
	ST_2	ST2		
Regular	pw_1	rOPI	cost1i = 35.5	$35.5 hourly per person
	m_p		mpi = 5	5 teams for each task
	h		hi = 30, 45, 50	Regular used time in respective tasks
	pw'_1	rOPJ	cost1j = 35.5	$35.5 hourly per person
	m_p		mpj = 5	5 teams for each task
	h		Hj = 30, 50, 60	Regular used time in respective tasks
Excess	pw_2	eOPI	cost2i = 50.5	$50.5 hourly per person
	pw'_2	eOPJ	cost2j = 50.5	$50.5 hourly per person
Insufficient	pw_3	iOPI	cost3i = 80.5	$80.5 hourly per person
	pw'_3	iOPJ	cost3j = 80.5	$80.5 hourly per person
Shift	c	c = 25, 25, 25, 25, 25, 25		$25 hourly per person in each shift
	s_{dt}	st = 34, 25, 13, 54, 36, 27		Used shift time in respective task
P	p_i	Demand	di = 150	Demand 150 people to finish the task
	p_j		dj = 150	Demand 150 people to finish the task
B		B	$\mathrm{B_i}$ = 1000	
			$\mathrm{B_j}$ = 1000	
x_{dti}		xi = $\left\{\begin{array}{l}1\\0\end{array}\right.$		Binary variable
x_{dtj}		xj = $\left\{\begin{array}{l}1\\0\end{array}\right.$		

And the hourly salary would be increased to $80.5 per person for insufficient manpower situation when insufficient manpower cases happen in festivals. Oppositely, compared with increased spread of insufficient manpower cases, hourly salary per person need to fluctuate down more since enterprise profits gradually drop. And the hourly salary of shift of $25 should be counted because that the workload during shift periods would fall in time interval between surplus and regular cases.

(b) *Setting of used time for distinct situations and shifts*

Each operator might meet various scenarios. Based on those various scenarios, used time for each task should have some discrepancies in elapsed time of each task.

Table 3 Variables

Variables	Coding	Description
v_{dtpi}	vi	Number of operators in each situation
v_{dtpj}	vj	
a_{tdi}	ai	
a_{tdj}	aj	
$E\sum_{i=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdi}(\theta)$	Ei	
$E\sum_{j=1}^{3}\sum_{d\in D}\sum_{t\in T}\alpha_{tdj}(\theta)$	Ej	
w_{dt}	w	

Considering the parallel sub-operations, bogie and body-shell assembly, two sub-operations should consume the equal time span to finish them. In this case, consuming time each task in respective sub-operation should be stable, given with 30, 50, and 60 h correspondingly.

Subsequently, a feasible solution of consuming shifts can be displayed in such numbers, 34, 25, 13, 54, 36, 27, to satisfy the requirements of finding optimal solutions of model.

3.2.2 Valuation of Variables

As can be seen from Table 4, manpower variables such as vi, vj, ai, aj, Ei, and Ej, preserve non-negative integers. Those non-negative integers can be shown as the feasibility of these outputs. Interestingly, the ai and aj equals to zero. This implies that insufficient situation would not happen under the conditions of model. It is noticeable that the reduced cost column in Table 4, can be considered as the amount of penalty. Developers would have to pay penalty to introduce one unit of specified

Table 4 Excerpted from report statement

Variable	Value	Reduced cost
vi (task 1/2/3)	26/26/26	5325/7987.5/8875
vj (task 1/2/3)	26/26/26	5325/8875/10,650
ai (task 1/2/3)	0/0/0	50.5/50.5/50.5
aj (task 1/2/3)	0/0/0	50.5/50.5/50.5
Ei (task 1/2/3)	20/20/20	0/0/0
Ej (task 1/2/3)	20/20/20	0/0/0
w (task 1/2/3/4/5/6)	5/4/2/7/5/4	850/625/325/ 135/900/675
xi (task 1/2/3)	1/1/1	0/0/0
xj (task 1/2/3)	1/1/1	0/0/0

Table 5 Excerpted from report statement

Row	Slack or surplus	Dual price
1	1,256,685	−1
2	454	0
39	0	−80.5

variable into the solution. For instance, reduced cost of shift manpower in task 1 has reduced cost of 850. Develops have to pay a penalty of 850 units to introduce the variable into the solution. In other words, the objective value would increase by 850 units for objective function of the minimization model in this paper. In this case, the marvelous reduced cost of variables is chiefly caused by deficiency of inputs. Subsequently, Table 4 has shown a comparison of the valuation for vi in distinct tasks. This output reveals that all tasks have the same operator's assignment. And the same operator's assignment implies that there is not huge difference in the working pace and ability for each operator. Eventually, what stands out in Table 4 is to ensure that there really exists the optimal balance between shifts and normal operation scheduling, which are shown by xi and xj.

3.2.3 Deviation of Constraints

Developers should distinguish the solutions to ensure if the model is really satisfied inequalities of the constraints. And the Slack or Surplus and Dual Price columns of report statement is used for distinguishing the solutions. Table 5 illustrates some of the main characteristics of the constraints.

The Row 1 represents the objective function, which has the objective value of 1,256,685 in second column, but dual price of (-1) would oppositely increase the objective value by 1 for minimization model of this paper. Apparently, compared with number of Row 39 in same column, Row 2 has larger deviation. Based on the implication of Slack or Surplus term, the previous comparison reveals that there are some errors in data inputs, which cause the divergence.

4 Conclusion

This research firstly gives out parallel operating plan by analyzing the assembly procedure for MTR. Thereby, this paper incorporates two sub-operations of assembly, three kinds of operator's assignment situation, two categories of manpower costs and several related operating constraints, to develop the model. This model can help in the effective management of the MTR maintenance manpower supply and shift schedule setting. In order to suitably modifying this model, the model is formulated as mixed integer linear programs. Especially, the shift cost is also considered in the objective function. And developer also constructs the constraints based on the practical requirements in MTR. Furthermore, the developer builds the model of this paper through Lingo. And the value of parameter is inputted by developer for experiment. With the report statement from Lingo, the paper continues to evaluate the application effect of model.

Compared with the past manpower supply planning models, this research has the following contributions: (1) It identified the importance of shift cost during manpower supply. Considering the shift cost for scheduling would accurately help MTR to minimize their maintenance cost. (2) This research developed parallel carriage assembly manpower supply plan to improve the effectiveness in different operator's assignment. This plan includes all operator's assignment situations: surplus, insufficient, and regular. (3) This research also successfully applied integer programming to develop a model in solving the maintenance manpower supply problems for assembly procedure. It has referential value for pervious procedures, disassembly and inspection procedure. (4) The results of experiment prove that proposed models and solution could be useful in actual operations.

Although the preliminary results show the feasibility of model, more testing and case studies still need to be conducted in the future. It should be mentioned that there does exists the deficiency of inputs in the numerical study of this paper. And this model might need other constraints for combining with the practical users. It will be better if there are comparisons with other models for manpower planning problem or there is data in real-life case to validate the model. Therefore, based on more reports, the researchers can statistically process data, so that users can grasp the limitations of model, before putting it into practical use. This could be a direction of future research. Eventually, the scope of this paper is confined to the assembly procedure for maintenance manpower supply plan procedure. How to improve effectiveness of those previous procedures, disassembly and inspection, could be a topic for future research [15].

Acknowledgements Authors would like to acknowledge the support from HK MTR maintenance team in Kowloon Bay Depot.

This work was supported in part by the Seed Fund for Basic Research in HKU under Grant 201906159001, and in part by HKU KE Impact Project Scheme under Grant KE-IP-2019/20-31.

References

1. Q.L. Xue, H.L. Zhai, R. Joshua, An urban island floating on the MTR station: a case study of the West Kowloon development in Hong Kong. Urban Des. Int. **15**, 191–207 (2010)
2. W.L. Hsu, Approximation algorithms for the assembly line crew scheduling problem. Math. Oper. Res. **9**, 376–383 (1984)
3. F. Croci, M. Perona, A. Pozzetti, Work force management in automated assembly systems (2000). Retrieved from https://www.inc.com/encyclopedia/assembly-line-methods.html
4. G. Zhang, Learning from the past, planning for the future. MTR Projects J. **5**, 45–48 (2015)
5. PRC Rail Consulting Ltd., The railway technical website. Retrieved from https://www.railway-technical.com/trains/rolling-stock-manufacture.html

6. K. Dziki, D. Krenczyk, Mixed-model assembly line balancing problem with tasks assignment, in *IOP Conference Series: Materials Science and Engineering* (2019)
7. M. Fischetti, From mixed-integer linear to mixed-integer bilevel linear programming, in *International Conference on Optimization and Decision Science*, Nov 2017
8. O.S. Ismaila, O.G. Akanbi, O.E. Charles-Owaba, Cost minimization approach to manpower planning in a manufacturing company. Pac. J. Sci. Technol. **10**, 191–196 (2009)
9. C.H. Chen, S.Y. Yan, M.J. Chen, Short-term manpower planning for MRT carriage maintenance under mixed deterministic and stochastic demands. Ann. Oper. Res. **181**, 67–88 (2010)
10. C.T. Zhang, Y. Gao, W.J. Li, L.X. Yang, Z.Y. Gao, Robust train scheduling problem with optimized maintenance planning on high-speed railway corridors: the China case. J. Adv. Transp. **1** (2018)
11. G. Karthikeyan, K.N. Krishnaswamy, Assembly manpower allocation under proportionality constraints. Eur. J. Oper. Res. **44**, 39–46 (1990)
12. P. Jaillet, G.G. Loke, M. Sim, Strategic manpower planning under uncertainty. Oper. Res. (2019)
13. F. Khoshniyat, J.T. Krasemann, Analysis of strengths and weaknesses of a MILP model for revising railway traffic timetables (2017)
14. R. Price, J. Hou, Implementing a statutory minimum wage in Hong Kong: appreciating international experiences but recognizing local Conditions. Common Law World Rev. **40**, 95–118 (2011)
15. X.J. Guo, M.H. Li, Dynamic task assignment algorithm based on multi-criteria. J. Comput. Appl. **28**, 2507–2509 (2009)

Relationships Between Human Capital, Organizational Integration, and Performance in Omni-Channel Retailing

Yixuan Han, Hanchen Li, Haochen Sun, and Guang Song

Abstract Several studies have identified that human resource management drives organizational integration in omni-channel (OC) retailing and is the key factor in determining the integration of OC supply chain. However, the relationship between human capital (HC), organizational integration, and firm performance has not been proved, and how HC affects firm performance under the specific aspects of organizational integration is unclear. This paper adopts the structural equation model (SEM) for empirical analysis and research and draws the conclusion that HC plays a key role in the integration of the OC organizational culture and organizational structure. Meanwhile, these three factors have a positive impact on firm performance. This study recommends optimization of OC enterprise management and provides inspiration for academic research in OC, supply chain, and human resource management.

Keywords Omni-channel · Human capital · Organizational integration

1 Introduction

A new approach to channel integration is developing as the distinction between online and offline channels is miniscule, the omni-channel (OC). The OC regards delivering a seamless customer experience between channels as its aim [1]. Besides, supply chain integration (SCI) is vital in achieving successful OC retailing [2]. Therefore, some researchers believed that SCI in OC retailing should be established from

Y. Han (✉) · H. Li · H. Sun · G. Song
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 17241035@bjtu.edu.cn

H. Li
e-mail: 17241037@bjtu.edu.cn

H. Sun
e-mail: 17241012@bjtu.edu.cn

G. Song
e-mail: songguang@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_50

three aspects—information, process, and organization—and effective integration can improve firm performance [3–5].

Although rapid growth in technology has boosted the demand for soft foundation, some researchers claimed that the ultimate solutions to SCI are human resource and organization strategies rather than advanced technology [6]. Similarly, current research has shown more than 60% impact of OC organizational integration in SCI [2]. Several scholars have also linked SCI with human resource management, especially in OC retailing, claiming that human capital (HC) is the main driver of SCI [7, 8].

According to the "resource-based view" (RBV), companies can gain competitive advantage by acquiring and controlling valuable, rare, irreplaceable, and inimitable resources, which include tangible and intangible assets [9, 10]. In supply chain management, RBV has been well established, and past studies have confirmed that integrating important resources in organizations has played a vital role in SCI [2]. Besides, some studies that have applied the dynamic RBV theory confirm the value of HC to the organization [11].

Therefore, this study empirically evaluates the relationships between HC, organizational integration, and performance and contributes to both literature and practice using the structural equation model (SEM). It also examine the impacts of integration strategy on the firm's performance, using the conceptual model developed under OC retailing.

The rest of this paper is organized as follows. Section 2 gives a literature review and the hypothesized conceptual model. Section 3 elaborates the empirical methods. Section 4 gives the result of research. Section 5 presents the conclusion and the discussion of the result.

2 Literature Review and Conceptual Model

2.1 Human Capital

Previous studies have mainly recommended the importance of HC in OC retailing through qualitative analysis, whereas there is little evidence through quantitative analysis [2, 12]. Some researchers indicated that in the channel integration, HC should be regarded as the strategic focus [12]. Moreover, a study claimed that OC retailers are supposed to redeploy human resources to enhance employee productivity and promote organizational integration [8]. Although there has been some advancement in the people's capabilities on the SCI in OC retailing, it remains understudied.

Although plenty of factors of HC have been examined, with few similar implications in the SCM literature, there has been no consensus yet. In this paper, HC is measured in three dimensions: skills [13], attitudes [14], and relationships [15]. Skills pertain to the depth and width of expertise, knowledge, and experience connected to job performance [14]. Attitudes relate to trust, commitment, and enthusiasm, as well as employees' engagement and dedication. In terms of relationships, as an significant resource of SCI, an individual's competence plays an important role in interpersonal relationships, because it promotes a culture of teamwork in problem solving [7].

Earlier literatures have paid attention to the relationship between HC and organizational integration [8]. However, little evidence sheds light on the impact of HC on organizational integration from the perspective of OC retailing.

2.2 Organizational Integration in SCI

The field of management has extensive studies on SCI, and academicians and enterprise practitioners have validated the positive impact of SCI on the economic benefits of an enterprise [16]. A study showed that OC organizational integration is a key determinant [2]. Furthermore, some scholars believed that organizational integration should be assessed from different perspectives, including organizational culture and organizational structure, which form a major part of the organization's leadership activity [17].

According to a few studies, organizational culture is "the set of shared, taken-for-granted implicit assumptions that a group holds and that determine how it perceives, thinks about and reacts to its various environments" [18]. Past literature showed that organizational culture contributes to SCI and enhances firm performance [19]. Therefore, organizations should encourage a series of values and then influence behavior and willingness to share knowledge [20].

Organizational structure is defined as a formal set of tasks, reporting relationships, and coordination systems. A study contended that the centralization of organizational structure is the underlying focus of SCI [21]. It was also proved by the research on Just In Time (JIT) and non-JIT companies, which showed major differences in some aspects of the structure [22]. In other words, JIT companies were far more integrated than non-JIT.

2.3 SCI and Performance

The relationship between SCI and performance has been examined in depth in literature since the twenty-first century. Both academia and the society have a relatively better understanding that a more advanced standard of SCI can improve enterprise operation, optimize financial indicators, and enhance customer satisfaction [23]. Extant studies believed that SCI helps enterprises to acquire resources and eliminate barriers between departments and enterprises. Therefore, it is positively related to improvement in the firm's financial efficiency and operational effectiveness [24].

A research revealed that organizational integration, which can be regarded as the primary task, is positively related to firm performance [8]. However, how HC affects firm performance under the specific aspects of organizational integration is not clear.

Therefore, this study aims to determine how HC and organizational integration contribute to the success of the firm performance.

2.4 Conceptual Model

Our research examines the relationships between HC, organizational integration, and firm performance under OC retailing. Figure 1 demonstrates the conceptual model, and the underlying concepts of the model are further expressed.

Considering the RBV, HC is the inimitable resources for firms to achieve competitive advantage, because their knowledge and interpersonal relationships are invaluable and exceptional [6]. A few studies have linked SCI with human resource management [7], claiming that HC is the main driver of SCI. Moreover, earlier studies also believe that HC is positively related to channel integration. Meanwhile, it has been shown that HC is a critical determinant of OC organizational integration [2]. Organizational integration should be measured by organizational culture and organizational structure [6]. For example, multi-skills and exceptional character, which are the key dimensions of HC, help in building a learning-oriented culture [14]. A motivated and trusted team contributes to establishing a cross-functional structure and strategic collaboration [14]. Therefore, the following hypotheses are proposed:

H1: HC in SCI is positively related to the organizational culture integration.
H2: HC in SCI is positively related to the organizational structure integration.

The RBV states that, a enterprise can enhance its relational capabilities to adjust supply chain activities among the departments and supply chain partners [25]. A more efficient organizational strategy including analysis, defensiveness, futurity, and proactiveness was found to be associated with higher performance [26], and internal integration was identified as the most important differentiator of firm performance. Meanwhile, earlier literature has confirmed the positive impact of SCI on enterprise economic benefits [16, 27].

The past study researched the relationship between organizational culture, innovation, and performance. Organizational culture was found to partially mediate the effects on innovation and organizational performance [28]. Besides, organizational structure was found to promote firm performance by influencing organizational effectiveness through routinized processes, tasks, and systems [29].

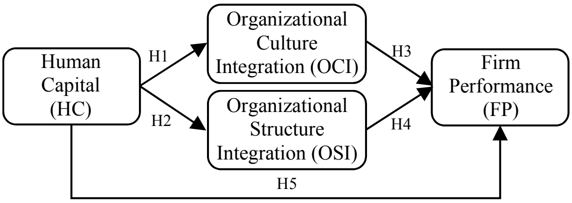

Human Capital (HC)
H1
H2
Organizational Culture Integration (OCI)
Organizational Structure Integration (OSI)
H3
H4
Firm Performance (FP)
H5

Fig. 1 Conceptual model

Moreover, organizational culture is a source of sustained competitive advantage [30]. Organizational structure is claimed to be a business resource that can be used to coordinate work that has been divided into smaller tasks [31]. Some studies suggested that organizational integration can employ the enterprise resources to improve the firm performance [24]. Therefore, the following hypotheses are proposed:

H3: Organizational culture integration in SCI is positively related to the firm performance.

H4: Organizational structure integration in SCI is positively related to the firm performance.

According to the RBV, social complexity resources and capabilities should be considered among the most important sources of sustainable competitive advantage for enterprises [32]. It is proposed that HC could result in competitive performance through transforming itself into competitive advantages [7]. The extant study states that a fit between firms' resources and environmental factors is essential to positively impact performance [33], and HC is considered as one of the most important intangible resources [34]. Therefore, the following hypothesis is proposed:

H5: HC is positively related to the firm performance.

3 Methodology

3.1 Sample Selection

The members of China General Chamber of Commerce (CGCC) are selected as the sample. In the questionnaire, the interviewed enterprises are first asked whether they have implemented the OC strategy. If so, the questionnaire will show specific questions to the enterprises; otherwise, the questionnaire will be suspended. A total of 1423 retailers were randomly selected from the database of CGCC for this study. After a thorough check, the responses gave 238 valid questionnaires with a response rate of 16.73%. Table 1 lists the overview of the respondents and their enterprises.

3.2 Questionnaire Design

Each factor is measured by a multi-item determined from the previous study, in order to make the proposed conceptual model operational. Similar to prior studies, the present paper has four factors, including HC, organizational culture integration, organizational structure integration, and firm performance (shown as Appendix).

Based on the conceptual model, the questionnaire contained three aspects: (1) the HC levels of OC retailing employees, (2) the internal and external integration, and (3) the operational and financial performance. The study utilized a five-point scale

Table 1 Overview of respondents and enterprises

Characteristics		Percentage (%)
Respondent's position	Top manager	51.26
	Middle manager	48.74
Ownership of the firm	State-owned	31.09
	Local private	50.84
	Foreign	7.14
	Joint venture	10.92
Age of the firm in China market (number of years)	Less than 5	13.03
	5–10	35.29
	11–15	24.37
	16–20	9.24
	More than 20	18.07
Annual revenue (million RMB Yuan)	Less than 50	24.37
	50–100	26.05
	100–200	20.59
	200–2000	20.59
	More than 2000	8.40

to determine what extent the respondents consent to each statement item (1 equals to particularly disagree or weak, 5 equals to particularly agree or strong).

4 Result

4.1 Preliminary Study

To test nonresponse bias and ensure the suitability of the collected data for the factor analysis, This study conducted the Kaiser–Meyer–Olkin (KMO) and Bartlett's test, in order to test nonresponse bias and make sure the collected data was suitable for the factor analysis. The result of the KMO test was 0.903. The Bartlett's test showed a content result when $p < 0.000$. Consequently, both results indicated that the data was appropriate for factor analysis (Table 2).

Table 2 KMO and Bartlett's test

Kaiser–Meyer–Olkin	Bartlett's test		
	χ^2	*df*	*Sig*
0.903	3,105.57	231	0

Table 3 Results of EFA

Factor	Number of items	Cronbach α	Mean	S.D
Human Capital	6	0.900	3.784	0.871
OCI	5	0.845	3.992	0.851
OSI	5	0.869	3.956	0.805
FP	6	0.895	3.773	0.806

4.2 Exploratory Analysis Results

.This paper first used Exploratory factor analysis (EFA) on 22 measured variables and extracted the critical factors by principal component analysis with VARIMAX rotation. Four factors were explained by these 22 variables (Table 3). The cumulative variance explained by the four factors was 0.630, and the alpha value was higher than 0.70 which is the suggested threshold in each group [35], indicating a good reliability (Table 3). Therefore, this result shows that the 22 variables can be explained by the four factors mentioned in the conceptual model.

4.3 Confirmatory Analysis Results

Confirmatory factor analysis (CFA) was used to examine the model fit. The result of the main-fit indices is adequate (χ^2 (203) = 384.859, RMSEA = 0.031, CFI = 0.939, GFI = 0.868, AGFI = 0.835, NFI = 0.880), indicating that the model, presented in Table 4, is completely acceptable. The *t*-value of all the items outdistances the critical ratio of a 0.05-level significance (Table 4). Besides, the R^2 value of each variable (>0.3) indicates abundant evidence of convergent validity [36]. Therefore, these 22 variables are significantly related to the factor construct. Furthermore, the t-value and R2 value results provided adequate evidence of convergent validity.

Further analysis can be run with assumption of an sufficient model fit, one of the effective analyses is the test of discriminant validity. To measure the discriminant validity, this paper built a constrained CFA model for each possible pair of latent constructs. Table 5 presents the results for 22 uncorrelated discriminant validity tests among the four latent factor groups. Under one degree of freedom, all χ^2 differences showed significant results that the discriminant validity was satisfied with a *p*-value less than 0.001.

The average variance extracted (AVE) was used as a further measurement to composite reliability. The results of four factor groups were all exceeding the critical value of 0.70 [36]. Thus, the analysis offered ample evidence of discriminant validity for the variables.

Table 4 Parameter estimates of CFA

Latent variable	Item	Standardized factor loading	*t*-value	R^2
HC	HC1	0.873		0.761
	HC2	0.860	15.422	0.739
	HC3	0.855	15.113	0.731
	HC4	0.839	14.996	0.705
	HC5	0.784	14.634	0.668
	HC6	0.793	11.249	0.692
OCI	OCI1	0.640		0.691
	OCI2	0.620	12.624	0.638
	OCI3	0.798	14.538	0.783
	OCI4	0.608	12.577	0.637
	OCI5	0.649	13.944	0.693
OSI	OSI1	0.727		0.693
	OSI2	0.694	13.003	0.645
	OSI3	0.637	12.398	0.589
	OSI4	0.647	12.488	0.600
	OSI5	0.668	9.670	0.623
FP	FP1	0.868		0.753
	FP2	0.849	19.430	0.721
	FP3	0.883	18.539	0.780
	FP4	0.866	20.210	0.749
	FP5	0.836	19.329	0.699
	FP6	0.894	17.965	0.799

Table 5 Discriminant validity

Measures	AVE	HC	OCI	OSI	FP
HC	0.672	1			
OCI	0.612	0.040	1		
OSI	0.610	0.021	0.827	1	
FP	0.750	0.664	−0.113	−0.160	1

4.4 Hypotheses Testing and Results

The hypotheses of all the four factors were tested by using a structural model. And the causality among the factors was evaluated by using the SEM with the maximum likelihood estimation method. Table 6 presents that the goodness of fit indices is $\chi^2 = 517.344$ with df = 204, RMSEA = 0.060, CFI = 0.936, NFI = 0.896, and TLI = 0.927. There is a effective indication of the model's fitness to the dataset.

Table 6 Results of the structural equation modeling

Variables	Estimate	S.E	C.R	P
HC → OCI	0.702	0.055	12.836	***
HC → OSI	0.786	0.065	12.033	***
OCI → FP	0.412	0.081	5.112	***
OSI → FP	0.177	0.060	2.969	**
HC → FP	0.165	0.087	2.216	**

After validating the fitness of the conceptual model, the hypothesized relationships are examined. Table 6 lists the results:

1. HC to Organizational Integration: H1 and H2 were supported, because the path coefficients are 0.702 ($t = 12.836$) and 0.786 ($t = 12.033$), respectively, which are both significant at the level of 0.001.
2. Organizational Integration to Firm Performance: H3 and H4 were supported, because the path coefficients are 0.412 ($t = 5.112$) and 0.177 ($t = 2.969$), which are significant at the level of 0.001 and 0.01, respectively.
3. HC to Firm Performance: H5 was supported, because the path coefficient is 0.165 ($t = 2.216$), which is significant at the level of 0.01.

The detailed assessment shows that HC and organizational integration are the critical factors that affect firm performance under the background of OC retailing. Therefore, retailers should concentrate on HC and organizational integration to improve the firm performance of OC retailing.

5 Conclusion and Discussion

Based on the RBV, this paper adopted SEM to assess the relationship between HC, organizational integration, and firm performance. The results showed that HC promotes organizational integration of OC retailing in terms of organizational culture and organizational structure. HC, organizational culture integration, and organizational structure integration all facilitate improvement of firm performance, among which promotion of organizational culture integration is the most significant, indicating that the integration of organizational culture driven by HC is one of the critical factors to improve performance. In addition, the effect of HC on firm performance is evident when organizational culture integration is used as the mediator variable.

5.1 Discussion About Results

The RBV suggests that firms are bundles of resources [10] including tangible and intangible assets [9]. The dynamic perspective of RBV leads some studies to suggest that HC leveraging can meet a firm's strategic need and require the co-specialized human-capital-oriented restructuring to promote a firm's flexible capabilities [11]. Besides, firms with superior HC outperform their competitors [37]. Therefore, this study examined how HC influences organizational integration on firm performance. Furthermore, SCI is an inimitable and nonsubstitutable resource based on RBV [38]. This study considered one of the most discussed logistics integration capabilities (organizational integration) to study its effect on firm performance.

The study results showed that the hypotheses of a positive relationship of HC on both organizational culture integration (H1) and organizational structure integration (H2) are supported. This result is in accordance with previous studies, which showed that employees contribute to the establishment of a culture and the construction of cross-functional structures and strategic collaboration. Therefore, this study indicates that HC can significantly improve integration of organizational culture and organizational structure to influence organizational integration. Besides, the results show that HC has a significant influence on the integration of organizational culture and organizational structure at the same confidence level ($p < 0.001$).

H3 and H4 are supported by the empirical results, which imply that organizational integration favors firm performance and that organizational culture integration has a greater impact on firm performance than organizational structure integration. This conclusion is consistent with that in the extant research, which stated that both organizational culture integration and organizational structure integration can enhance a firm's performance. In addition, organizational culture integration in organizational integration is the main factor that has a significant positive impact on firm performance, while the effect of organizational structure integration is weaker than that of organizational culture integration. This confirms the findings in prior literature, which indicates that organizational culture has a vital influence on organizational performance improvement [29].

The results show that the positive relationship between HC and firm performance, given by hypothesis, is supported. This is consistent with the conclusions obtained in the past studies. Furthermore, HC significantly positive influences firm performance at the 0.01 confidence level, whereas the empirical results support H5, which implies a positive effect of HC on enterprise's performance and shows that there is a greater influence when retailers adopt an organizational culture integration strategy. Therefore, organizational culture integration is used as a mediator variable to improve the effect of HC on firm performance.

5.2 Implications for Research

This paper examines integration of organizational culture and organizational structure in the context of OC retailing using organizational variables. The study examined 238 retailers in China, and 22 measured items were selected according to the previous literature and categorized into four groups to develop the conceptual model. In OC retailers, the structure of the enterprise is more complex and requires multi-skilled talents [39]. The integration via this model can optimize the enterprise and significantly improve enterprise performance. However, the single channel retailing has a single structure, and the integration will not have a more obvious effect on the performance of single channel retailers.

First, the findings are accordance with the findings of the previous studies that showed HC to favor integration of organizational culture and organizational structure. Moreover, this paper states that integration of organizational culture and organizational structure favors firm performance, especially organizational culture integration. Second, this study presents that retailers' firm performance can improve significantly through HC when organizational culture integration is used as a mediator variable. This study figures out the positive effect of HC on firm performance as proposed in earlier studies and identifies the role of organizational culture integration in linking HC and firm performance.

5.3 Implications for Practice

The study provides valuable implications for practitioners to implement HC strategy, organizational culture integration, and organizational structure integration strategy. First, retailers are supposed to focus on organizational integration, especially on organizational culture integration strategy, because it plays a dominant role in improving the firm performance. Therefore, OC retailing should emphasize organizational culture integration, such that internal and external organizations are in a harmonious, cooperative, and communicative atmosphere. Second, OC retailing should apply organizational integration through HC. Organizational integration in OC retailing depends on enhancing the quality of employees. In an environment of openness, trust, and communication, HC-driven organizational integration significantly helps in promoting firm's performance.

5.4 Limitation

Although this study revealed some valuable findings, there are a few limitations that help identify issues needing further studies. First, the data merely comes from the retailer in China's market. Thus, the finding is not generalizable. Second, the questionnaire adopts Likert five-point scale design method, and the responses are mainly based on subjective judgment, which may lead to some deviation of the collected data. Third, due to the complex market environment, more control variables must be allowed for.

Appendix

List of measured items

Factor	Measured variable	Item
Human Capital (HC)	Employees have excellent post skills	HC1
	Employees have rich practical experience	HC2
	Employees are fully responsible and willing to provide suggestions for the enterprise	HC3
	Employees are honest and enthusiastic	HC4
	Employees are willing to cooperate with colleagues from other departments	HC5
	Employees and partners have a good relationship	HC6
Organization Culture Integration (OCI)	Enterprise culture is characterized by trust and openness	OCI1
	Enterprise encourages teamwork and interdepartmental collaboration	OCI2
	Enterprise encourages collaboration with supply chain partners	OCI3
	There is a consistent set of values and goals across the enterprise's channels	OCI4
	Enterprise understands and respects the enterprise culture of its supply chain partners	OCI5
Organization Structure Integration (OSI)	The organizational structure facilitates the exchange of information between departments	OSI1
	Enterprise organizes interdepartmental and interchannel meetings for internal communication regularly	OSI2
	The organizational structure facilitates the flow of information between supply chain partners	OSI3
	Enterprise has special communication mechanisms with their supply chain partners	OSI4

(continued)

(continued)

Factor	Measured variable	Item
	Enterprise has a dedicated team to communicate with supply chain partners	OSI5
Firm Performance (FP)	Improvement of service level	FP1
	Improvement in speed response	FP2
	Increase of market share	FP3
	Rise of revenue	FP4
	Reduction of operations cost	FP5
	Improvement of return on assets	FP6

References

1. S. Gallino, A. Moreno, Integration of online and offline channels in retail: the impact of sharing reliable inventory availability information. Manage. Sci. **60**(6), 1434–1451 (2014)
2. G. Song, S. Song, L. Sun, Supply chain integration in omni-channel retailing: a logistics perspective. Int. J. Logist. Manag. **30**(2), 527–548 (2019)
3. N. Beck, D. Rygl, Categorization of multiple channel retailing in multi-, cross-, and omni-channel retailing for retailers and retailing. J. Retail. Consum. Serv. **27**, 170–178 (2015)
4. D.M. Russell, A.M. Hoag, People and information technology in the supply chain: social and organizational influences on adoption. Int. J. Phys. Distrib. Logist. Manag. **34**(2), 102–122 (2004)
5. S. Chopra, How omni-channel can be the future of retailing. DECISION **43**(2), 135–144 (2016)
6. A.N. Shub, P.W. Stonebraker, The human impact on supply chains: evaluating the importance of ‘soft’ areas on integration and performance. Supply Chain Manag. **14**(1), 31–40 (2009)
7. Z. Wang, B. Huo, Y. Qi, X. Zhao, A resource-based view on enablers of supplier integration: evidence from China. Ind. Manag. Data Syst. **116**(3), 416–444 (2016)
8. S. Song, X. Shi, G. Song, Supply chain integration in omni-channel retailing: a human resource management perspective. Int. J. Phys. Distrib. Logist. Manag. **50**(1), 101–121 (2019)
9. D. Xu, B. Huo, L. Sun, Relationships between intra-organizational resources, supply chain integration and business performance: an extended resource-based view. Ind. Manag. Data Syst. **114**(8), 1186–1206 (2014)
10. B. Wernerfelt, Harmonised implementation of application-specific messages (ASMs). Strateg. Manag. J. **CINCO**(2), 1–12 (1984)
11. C.Y.P. Wang, B.S. Jaw, C.H.C. Tsai, Building dynamic strategic capabilities: a human capital perspective. Int. J. Hum. Resour. Manag. **23**(6), 1129–1157 (2012)
12. M. González-Loureiro, M. Dabic, F. Puig, Global organizations and supply chain: new research avenues in the international human resource management. Int. J. Phys. Distrib. Logist. Manag. **44**(8–9), 689–712 (2014)
13. Z. Cao, B. Huo, Y. Li, X. Zhao, The impact of organizational culture on supply chain integration: a contingency and configuration approach. Supply Chain Manag. **20**(1), 24–41 (2015)
14. B. Huo, Y. Ye, X. Zhao, Y. Shou, The impact of human capital on supply chain integration and competitive performance. Int. J. Prod. Econ. **178**, 132–143 (2016)
15. Y. Shou, Y. Li, Y. Park, M. Kang, Supply chain integration and operational performance: the contingency effects of production systems. J. Purch. Supply Manag. **24**(4), 352–360 (2018)
16. L. Du, Acquiring competitive advantage in industry through supply chain integration: a case study of Yue Yuen Industrial Holdings Ltd. J. Enterp. Inf. Manag. **20**(5), 527–543 (2007)

17. M. Russo, M. Cesarani, Strategic alliance success factors: a literature review on alliance lifecycle. Int. J. Bus. Adm. **8**(3), 1 (2017)
18. E.H. Schein, Organizational culture and leadership: a dynamic view. Organ. Stud. **7**, 199–201 (1985)
19. S. Mullarkey, P.R. Jackson, S.K. Parker, Employee reactions to JIT manufacturing practices: a two-phase investigation. Int. J. Oper. Prod. Manag. **15**(11), 62 (1995)
20. Y. Li, M. Tarafdar, S.S. Rao, Collaborative knowledge management practices: theoretical development and empirical analysis. Int. J. Oper. Prod. Manag. **32**(4), 398–422 (2012)
21. T.P. Stank, P.J. Daugherty, C.M. Gustin, Organizational structure: influence on logistics integration, costs, and information system performance. Int. J. Logist. Manag. **5**(2), 41–52 (1994)
22. R. Germain, C. Droge, The context, organizational design, and performance of JIT buying versus non-JIT buying firms. Int. J. Purch. Mater. Manag. **34**(1), 12–18 (1998)
23. D. Lu, Y. Ding, S. Asian, S.K. Paul, From supply chain integration to operational performance: the moderating effect of market uncertainty. Glob. J. Flex. Syst. Manag. **19**, 3–20 (2018)
24. G. Willis, S.E. Genchev, H. Chen, Supply chain learning, integration, and flexibility performance: an empirical study in India. Int. J. Logist. Manag. **27**(3), 755–769 (2016)
25. J. Lewis, P. Whysall, C. Foster, Drivers and technology-related obstacles in moving to multichannel retailing. Int. J. Electron. Commer. **18**(4), 43–67 (2014)
26. F. Bergeron, L. Raymond, S. Rivard, Ideal patterns of strategic alignment and business performance. Inf. Manag. **41**(8), 1003–1020 (2004)
27. R. Narasimhan, S.W. Kim, Effect of supply chain integration on the relationship between diversification and performance: evidence from Japanese and Korean firms. J. Oper. Manag. **20**(3), 303–323 (2002)
28. K. Abdi et al., The effect of knowledge management, organizational culture and organizational learning on innovation in automotive industry. J. Bus. Econ. Manag. **19**(1), 1–19 (2018)
29. W. Zheng, B. Yang, G.N. McLean, Linking organizational culture, structure, strategy, and organizational effectiveness: mediating role of knowledge management. J. Bus. Res. **63**(7), 763–771 (2010)
30. J. Barney, Firm resources and sustained competitive advantage. J. Manage. **17**(1), 99–120 (1991)
31. P.J. Daugherty, H. Chen, B.G. Ferrin, Organizational structure and logistics service innovation. Int. J. Logist. Manag. **22**(1), 26–51 (2011)
32. J.B. Barney, P.M. Wright, On becoming a strategic partner: the role of human resources in gaining competitive advantage. Hum. Resour. Manage. **37**(1), 31–46 (1998)
33. D.C. Hambrick, D. Lei, Toward an empirical prioritization of contingency variables for business strategy. Acad. Manag. J. **28**(4), 763–788 (1985)
34. D.I. Prajogo, The strategic fit between innovation strategies and business environment in delivering business performance. Int. J. Prod. Econ. **171**(SI), 241–249 (2016)
35. J.B. Skipper, J.B. Hanna, Minimizing supply chain disruption risk through enhanced flexibility. Int. J. Phys. Distrib. Logist. Manag. **39**(5), 404–427 (2009)
36. J. Hair, R. Anderson, R. Tatham, W. Black, *Multivariate Data Analysis* (Prentice Hall, Upper Saddle River, NJ, 1998)
37. P.M. Wright, B.B. Dunford, S.A. Snell, Human resources and the resource based view of the firm. J. Manage. **27**(6), 701–721 (2001)
38. J.B. Barney, Purchasing, supply chain management and sustained competitive advantage: the relevance of resource-based theory. J. Supply Chain Manag. **48**(2), 3–6 (2012)
39. P.C. Verhoef, P.K. Kannan, J.J. Inman, From multi-channel retailing to omni-channel retailing. Introduction to the special issue on multi-channel retailing. J. Retail. **91**(2), 174–181 (2015)

Research on the Cashback Strategy of E-tailers with Strategic Consumers

Danqin Yang, Nan Yan, and Jiani Ding

Abstract We integrate the cashback promotion strategy and the strategic consumer's waiting behavior into a framework and consider a two-stage model in which an e-tailer sells products to strategic customers through two online platforms, e-shop and cashback website (CW). We develop a model to characterize the equilibrium decisions for each participant and investigate the impacts of cashback on the decisions and profits of e-tailer and customers. We found e-tailers can restrain strategic consumer behavior by reasonable use of cashback websites. And e-tailers can make the CW plays the role of advertising and price discrimination through pricing. When the CW only plays the role of advertising, e-tailers can make profits from it.

Keywords Cashback website · Strategic consumer · Rational expectation equilibrium · Newsvendor model

1 Introduction

The nature of a cashback website is a promotional platform based on sales. It has been widely used in the field of e-commerce and has become a kind of new internet service. Through cooperation with cashback website platform, e-tailers promote their products, attract customers with rebates, and expand product sales, so as to occupy the market and gain profits. It acts as an intermediary between merchants and customers, enabling e-tailers to gain more profits, customers to obtain greater discounts, and itself to earn a certain commission, thus achieving a win-win-win situation.

Major e-commerce platforms have launched various promotional activities, such as T-mall 618, "Double 11" Shopping Carnival and so on. Promotions frequently lead to the price changes frequently. Thus, smart consumers will compare the current price

D. Yang (✉) · N. Yan · J. Ding
Nanjing University of Science & Technology, School of Economics and Management, Nanjing, China
e-mail: yangdanqin@163.com

N. Yan
e-mail: 18251901099@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_51

with the expected promotion price in the future, and might wait for discount. This type of consumers is known as strategic consumers. The presence of strategic consumers makes the decisions of merchants more complicated. Some scholars have shown that ignoring strategic consumer behaviors will lead to 30% profit loss [1].

Although the importance of understanding and analyzing strategic consumer behavior is widely recognized, to the best of our knowledge, few studies have studied its impact on supply chain management under cashback mechanism. Moreover, few articles combine strategic consumer behavior with CWs to study the interaction between them. In view of this literature gap, we develop a model framework to study the strategic consumer behavior with cashback mechanism aimed at exploring a new way to deal with strategic consumer behavior.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related literature. In Sect. 3, we present the model and its mathematical formulation in detail. In Sect. 4, a numerical example is used to verify the model. We conclude the paper in Sect. 5.

2 Literature Review

The literature related to this paper mainly comes from three streams, i.e., consumer rebate, cashback website and strategic consumer behavior.

One of the related streams of research is consumer rebate. The field of marketing management focuses on the impact of rebate on consumer behaviors. Reference [2] show that rebate can stimulate consumption in current stage. Reference [3] verified that rebates can induce customers to give favorable comments, so as to influence the behavior of other consumers and improve the profits. The field of operation management focuses on the impact of rebate on supply chain. For example, [4] prove that rebates can help manufacturers manage inventory. Reference [5] proving that rebates can help manufacturers suppress the gray market. Reference [6] studied the rebate offered by manufacturers to customers in the two-level supply chain, and found that when customers are heterogeneous, rebates to customers make manufacturers profitable. Reference [7] compared three kinds of consumer rebate modes in the case of a single manufacturer, and proved that all kinds of rebate modes are profitable.

Another related stream of research is cashback websites. References [8, 9] use the data from cashback platforms to study the consumer satisfaction of cashback websites and the impact on consumer behavior. Reference [10] verified through numerical experiments that cashback websites can increase the possibility and scale of purchase by consumers. Some scholars pay attention to the pricing and effectiveness of cashback websites when they exist. For example, [11] compared the charging modes of two cashback websites, “conversion fee” and “guided fee”, and illustrated the applicability of the two modes. Reference [12] discovered the “cashback paradox”, proving that consumers may face a higher price in the presence of cashback websites. Reference [13] found that retailers could make profits by using cashback websites in both the centralized and decentralized supply chain settings.

The last related stream of research is strategic consumer behavior. The waiting behavior of strategic consumers makes the decision-making of merchants more complicated. How to deal with strategic consumers has always been the focus of scholars. Reference [14] pointed out that providing price guarantee is conducive to eliminating consumers' strategic behaviors. Reference [15] implemented two mechanisms, namely price commitment and inventory commitment, to eliminate the impact of customer strategic behavior. Reference [16] found that in a decentralized supply chain, price and inventory commitments can exacerbate the double marginalization inefficiency of the supply chain.

The two literatures closely related to this study are [13] and [15]. The former studied a single-cycle decision-making problem of e-shop with cashback websites without considering strategic consumer behavior. Different from this research, this paper introduces strategic consumers and establishes a two-period model. The latter studied the effects of quantity commitment and price commitment on strategic consumer behavior, and studied the performance of the two kinds of commitments in the markdown money contract and the rebate sales contract. However, they studied rebates from manufacturers to retailers, while this paper focuses on rebates from e-tailers to consumers via cashback websites.

3 Model

A. *Model Description*

We consider a two-stage newsvendor model. An e-tailer sells merchandise via e-shop and Cashback Website (CW). The e-shop sells products to consumers directly and the CW provides the e-shop with a link service. Consumers can purchase merchandise from the e-shop by visiting e-shop directly or following a link from the CW. In the real e-retail scenario, the cashback website can be divided into two types according to the website ownership. One is the third-party cashback platform authorized by the e-shop, such as fanli.com. The other is self-owned cashback platform. That is the cashback website and the e-shop belong to the same entity, such as Taobao.com and Etao.com, QQ.com and its affiliated QQ rebate. For the former, it collects a percentage of the commission from the e-tailer for each transaction that takes place through the link to the site and returns a portion of the commission to the consumer as a cashback. For the latter, we can assume that the company that owns the e-shop and CW can make the best decision through internal coordination. Thus, we can omit the decision of commission, and only decide the cashback for consumers. In this paper, we consider the e-tailer's optimal decision problem under third-party cashback website which is a more general case.

The selling season is divided into a full-price period (selling in original price p) and a clearance period (selling in promotion price s). Assume that the e-tailer sells a single product over two periods, and has only one opportunity to place an order, that it occurs before the start of the selling season. We do not consider the loss of stock out and the disposal cost of overstock, and assumes $p > c > s$, so that the marginal revenue of the e-tailer is positive in the full-price period, and he can clear the inventory in the clearance period. In addition, we assume that the CW only provides a cash bonus in the first selling period.

Suppose that the e-tailer faces the random demand σ, which can be regarded as the total number of consumers in the market. Let $F(\cdot)$ and $f(\cdot)$ be the cumulative distribution and density function of σ, respectively. We assume $f(\cdot)$ is continuous and $f(0) > 0$.

We assume that all consumers in the market are homogeneous strategic consumers and have the same valuation v on the product and do not decrease over time. Each consumer only buys one unit of products. In reality, due to factors such as the reputation of the cashback website, consumers may distrust the quality, brand authorization and other aspects of products linked by the website, resulting in the reduction of consumer valuation. So that we use the $\theta(0 < \theta < 1)$ as a discount factor to characterize the reduction in consumer valuation caused by the use of cashback website.

The sequence of decisions is as follows. First, the e-tailer forms beliefs ξ_r over consumers' reservation prices and then optimally chooses the price p and quantity Q. Next the customers privately form beliefs ε over their chances of obtaining the product in the salvage market and form their reservation prices r based on these beliefs, and choose between buying right now and waiting. Then the random demand σ is realized. Finally, all remaining units are sold at the salvage price s. Moreover, consumers can only observe the retailer's price when making decisions, not its inventory.

B. *The Benchmark Model Without Cashback Website*

We develop a benchmark model without cashback. The retailer sells his product only via e-shop. The sequence of decisions are as that in model description. Before selling season, the e-tailer forms beliefs ξ_r over consumers' reservation prices and then set the price $p = \xi_r$, and quantity $Q = arg_Q \max \prod(p)$. Next the consumers privately form beliefs ε over their chances of obtaining the product on the salvage market and form their reservation prices r based on these beliefs. The utility obtained by the strategic consumers in the full-price period is $v - p$, and the expected utility be obtained by the strategic consumer in the clearance period is $(v - s)\varepsilon$. So the strategic consumers is willing to buy in the full-price period when $v - p \geq (v - s)\varepsilon$. Thus, the strategic consumers' reservation price is $r = v - (v - s)\varepsilon$. Thus, if $p < r$, the strategic consumers choose buying right now; otherwise, waiting for the discount. We find the rational expectation (RE) equilibrium for the above problem. We define the equilibrium as follows.

Defination 1 *A RE equilibrium* $(p, Q, \xi_r, r, \varepsilon)$ *satisfies the following conditions*:
(i) $r = v - (v - s)\varepsilon$, (ii) $p = \xi_r$, (iii) $\xi_r = r$, (iv) $Q = arg_Q \max \prod (Q, p)$, (v) $\varepsilon = F(Q)$.

According RE equilibrium conditions, the above problems can be simplified as

$$p = v - (v - s)F(Q) \tag{1}$$

Let’s move on to retailer. The retailer’s profit function is

$$\prod = p\mathrm{E}[\min\{\sigma, Q\}] + s(Q - \mathrm{E}[\min\{\sigma, Q\}]) - cQ \tag{2}$$

where E(·) represents expectation operation. In Eq. (2), the first item is the sales revenue in full-price period. The second item is the sales revenue in clearance period. The third item is the ordering cost. Solving the e-tailer’s optimization problem, we get Eq. (3), where, we use $\bar{F}$ to denote complementary cdf $1 - F$.

$$\bar{F}(Q) = \frac{c - s}{p - s} \tag{3}$$

From Eqs. (1)–(3), we derive Proposition 1. Note that we use the superscript ‘*’ to mark the equilibrium solutions in the remainder of this article.

Proposition 1

(a) *In the RE equilibrium, all strategic consumers buy immediately, and the e-tailer’s price and quantity are characterized by*

$$p^* = s + \sqrt{(c - s)(v - s)}, \quad \bar{F}\left(Q^*\right) = \sqrt{\frac{c - s}{v - s}} \tag{4}$$

(b) *The optimal profit of the e-tailer can be expressed as*

$$\prod\left(Q^*, p^*\right) = \sqrt{(c - s)(v - s)} * \int\limits_{0}^{\bar{F}^{-1}\left(\sqrt{\frac{c-s}{v-s}}\right)} xf(x)dx \tag{5}$$

Proofs of all propositions are given in Appendix. We can easily compare the equilibrium quantity and price in our model with the newsvendor model, where customers are not strategic and the retailer set price equal to consumer valuation *v*. The equilibrium price p^* must lower than *v*, because $p^* - s = \sqrt{(c - s)(v - s)}$. It shows that the retailer has to lower price below *v* to deal with strategic consumer. In terms of quantity, the equilibrium stocking quantity Q^* is also lower than Q_0, the equilibrium stocking quantity of standard newsvendor model without strategic consumers, because $\bar{F}(Q_0) = \frac{c-s}{v-s} < \sqrt{\frac{c-s}{v-s}} = \bar{F}(Q^*)$. This is a common tactic used by retailers to increase customers’ willingness to pay by restricting the availability of the product.

C. *The Model With Third-party Cashback Website*

We assume that the e-tailer has a contractual relationship with the third-party CW, and the single e-tailer sells his product via e-shop and CW. For products selling through the links provided by the CW, the e-tailer shall pay a commission of βp *t*o the CW, and the cashback website shall return αp to the consumers as a reward, where $0 < \beta < 1, 0 < \alpha < \beta$. From the existing research, we know that the cashback website, as a means of promotion, can help e-tailers to implement three levels of price discrimination on one hand [12]. On the other hand, it can expand consumers demand [10]. We use $D(\alpha, \sigma)$ to represent the market demand function when using the CW, where $g(\cdot)$ and $G(\cdot)$ are respectively the density function and the cumulative distribution function, and σ represents the random demand which is not affected by cashback.

According to the RE equilibrium condition, consumers' estimation of product availability in the second sales period is $\varepsilon = G(Q|\alpha)$. Therefore, the strategic consumer's decision is as follows. If strategic consumers buy through e-shop, the utility obtained by the strategic consumers in the full-price period is $v - p$. If strategic consumers buy through CW, the utility may be obtained by the strategic consumer in the full-price period is $\theta v - (1 - \alpha)p$. If strategic consumers choose waiting, their utility is $(v - s)G(Q|\alpha)$. Therefore, the consumers' reservation price is as follows:

(1) When $v - p \geq \theta v - (1 - \alpha)p$, i.e., $v \geq \frac{\alpha p}{1-\theta}$, the strategic consumers choose buy through e-shop. And their reservation price is $r = v - (v - s)\varepsilon$, and if $p \leq r$, the strategic consumers choose buying right now, otherwise, waiting for discount. In this case, based on the RE equilibrium conditions, we can get

$$p = v - (v - s)G(Q|\alpha) \tag{6}$$

(2) When $\theta v - (1 - \alpha)p > v - p$, i.e., $v < \frac{\alpha p}{1-\theta}$, the strategic consumers choose to buy through CW. And their reservation price is $r = \frac{\theta v - (v-s)\varepsilon}{1-\alpha}$, and if $p \leq r$, the strategic consumers choose buying right now, otherwise, waiting for discount. And similarly, based on the RE equilibrium conditions, we can get

$$p = [\theta v - (v - s)G(Q|\alpha)]/(1 - \alpha) \tag{7}$$

Let's move to the e-tailer's profit. According to the above analysis, the e-tailer's profit function is also divided into the following two cases.

(1) When $v \geq \frac{\alpha p}{1-\theta}$, the strategic consumers choose buy through e-shop, and the e-tailer's profit function is

$$\prod_{es}(Q, p) = (p - s)E[\min(D(\alpha, \sigma), Q)] - (c - s)Q \tag{8}$$

(2) When $v < \frac{\alpha p}{1-\theta}$, the strategic consumers choose to buy through CW, and the e-tailer's profit function is.

$$\prod_{cw}(Q, p) = ((1-\beta)p - s)E[\min(D(\alpha,\sigma), Q)] - (c-s)Q \tag{9}$$

Here we use subscript "es" to represent "e-shop", subscript "cw" to represent "CW".

Next we consider the e-tailer's optimal decisions under two different forms of market demand, namely additive demand and multiplicative demand. In the form of additive demand, the expression of the requirement function is $D(\alpha,\sigma) = d(\alpha)+\sigma$, $d(\alpha)$ represents the demand function which is affected by cashback. Without loss of generality, we assume that $d'(\alpha) > 0, d''(\alpha) \le 0$. We can easily derive $G(Q|\alpha) = F(Q - d(\alpha))$. Similarly, in the form of multiplicative demand, the expression of the demand function is $D(\alpha,\sigma) = d(\alpha)\sigma$, and $G(Q|\alpha) = F(Q/d(\alpha))$.

- The model under the additive demand
  The e-tailer's expected demand function under additive demand is

$$E[\min(D(\alpha,\sigma), Q)] = \int_{0}^{Q-d(\alpha)} (x + d(\alpha))f(x)dx + \int_{Q-d(\alpha)}^{\infty} Qf(x)dx \tag{10}$$

Substituting (10) into (8) and (9), we have

$$\begin{aligned}\prod_{es} &= (p-s)\left[\int_{0}^{Q-d(\alpha)} (x + d(\alpha))f(x)dx + \int_{Q-d(\alpha)}^{\infty} Qf(x)dx\right] - (c-s)Q \\ \prod_{cw} &= ((1-\beta)p-s)\left[\int_{0}^{Q-d(\alpha)} (x + d(\alpha))f(x)dx + \int_{Q-d(\alpha)}^{\infty} Qf(x)dx\right] - (c-s)Q\end{aligned} \tag{11}$$

Solving these problems, then we derive Proposition 2.

Proposition 2 *With additive demand, we have*

(a) *When $v \ge \frac{\alpha p^*}{1-\theta}$, the e-tailer's optimal price and optimal order quantity are characterized by*

$$p^* = s + \sqrt{(c-s)(v-s)}$$

$$Q^* = \bar{F}^{-1}\left(\sqrt{\frac{c-s}{v-s}}\right) + d(\alpha) \tag{12}$$

(b) *When* $v < \frac{\alpha p^*}{1-\theta}$*, the e-tailer's optimal price and optimal order quantity are characterized by*

$$\begin{aligned} p^* &= \frac{-\gamma + \sqrt{\gamma^2 - \delta}}{2(1-\alpha)(1-\beta)} \\ Q^* &= \bar{F}^{-1}\left(\frac{-\gamma + \sqrt{\gamma^2 - \delta}}{2(1-\alpha)(1-\beta)}\right) + d(\alpha) \end{aligned} \tag{13}$$

where $\gamma = (1-\beta)(1-\theta)v - (2-\alpha-\beta)s$,

$$\delta = 4(1-\alpha)(1-\beta)[s\theta v - c(v-s)].$$

- The model under the multiplicative demand
  The e-tailer's expected demand function under multiplicative demand is

$$E[\min(D(\alpha,\sigma),Q)] = \int_0^{\frac{Q}{d(\alpha)}} x d(\alpha) f(x)dx + \int_{\frac{Q}{d(\alpha)}}^{\infty} Qf(x)dx \tag{14}$$

Similarly, substituting (14) into (8) and (9), we have

$$\begin{aligned} \prod_{es} &= (p-s)\left[\int_0^{\frac{Q}{d(\alpha)}} x d(\alpha) f(x)dx + \int_{\frac{Q}{d(\alpha)}}^{\infty} Qf(x)dx\right] - (c-s)Q \\ \prod_{cw} &= ((1-\beta)p-s)\left[\int_0^{\frac{Q}{d(\alpha)}} x d(\alpha) f(x)dx + \int_{\frac{Q}{d(\alpha)}}^{\infty} Qf(x)dx\right] - (c-s)Q \end{aligned} \tag{15}$$

Solving these problems, then we derive Proposition 3.

Proposition 3 *ith multiplicative demand, we have*

(a) *When* $v \geq \frac{\alpha p^*}{1-\theta}$*, the e-tailer's optimal price and optimal order quantity are characterized by*

$$\begin{aligned} p^* &= s + \sqrt{(c-s)(v-s)} \\ Q^* &= \bar{F}^{-1}\left(\sqrt{\tfrac{c-s}{v-s}}\right) d(\alpha) \end{aligned} \tag{16}$$

(b) *When $v < \frac{\alpha p^*}{1-\theta}$, the e-tailer's optimal price and optimal order quantity are characterized by*

$$\begin{aligned} p^* &= \frac{-\gamma+\sqrt{\gamma^2-\delta}}{2(1-\alpha)(1-\beta)} \\ Q^* &= \bar{F}^{-1}\left(\frac{-\gamma+\sqrt{\gamma^2-\delta}}{2(1-\alpha)(1-\beta)}\right)d(\alpha) \end{aligned} \tag{17}$$

Comparing Propositions 1, 2 and 3, we find that, the equilibrium price of the e-tailer is the same with additive and multiplicative demand. When $v \geq \frac{\alpha p^*}{1-\theta}$, CW plays an advertising role. That means the use of CW doesn't affect the equilibrium price, but only affects the equilibrium quantity. When $v < \frac{\alpha p^*}{1-\theta}$, the CW plays the role of advertising and price discrimination. That is it affects both the equilibrium price and the equilibrium quantity.

4 Numerical Analysis

In this section, we conduct numerical examples to complement our analytical findings in the previous section. We assume that the demand σ is uniformly distributed over (500, 1000). Then we can easily know that $f(x) = 1/500$, and $F(Q) = (Q - 500)/500$. The random demand function affected by the cashback is $d(\alpha) = 200\alpha$. The default values of all parameters are given as: $\beta = 0.2, c = 200, s = 100, v = 600, \theta = 0.9$. From Figs. 1, 2 and 3, we derive the following observations.

Figure 1 shows the change of equilibrium price with α. From the previous model analysis, we can find that when an e-tailer sells products through a CW, his optimal price is only two cases. If the CW only plays the role of advertising, i.e., $v \geq \frac{\alpha p^*}{1-\theta}$, the e-tailer's optimal price is the same as the optimal price without the CW. Moreover, if the CW plays both roles of advertising and price discrimination, i.e., $v < \frac{\alpha p^*}{1-\theta}$, the e-tailer's optimal price increase with the cash back ratio, and is always higher than that without the CW. This reflects that e-tailers with third-party CW provide a high cash back to create an illusion of discount to consumers. That means cash back make consumers have to face a higher original price.

Figures 2 and 3 respectively show the effect of α on the equilibrium quantity and equilibrium profit. In order to express the conclusions in different cases more clearly, we partially enlarged Fig. 2a to get Figs. 2b, and 3a to get Fig. 3b. Obviously, in Fig. 2a, with the increase of the cashback ratio, the e-tailer's inventory increases, indicating that the cashback increases the demand of consumers, which reflects from the side that the CW can inhibit the strategic behavior of consumers and encourage consumption. In Fig. 2b, we can see when the cashback ratio is small, the quantity with CW is lower than that without CW. With the increase of cashback ratio, the quantity with CW will continue to increase, and eventually be higher than the order quantity

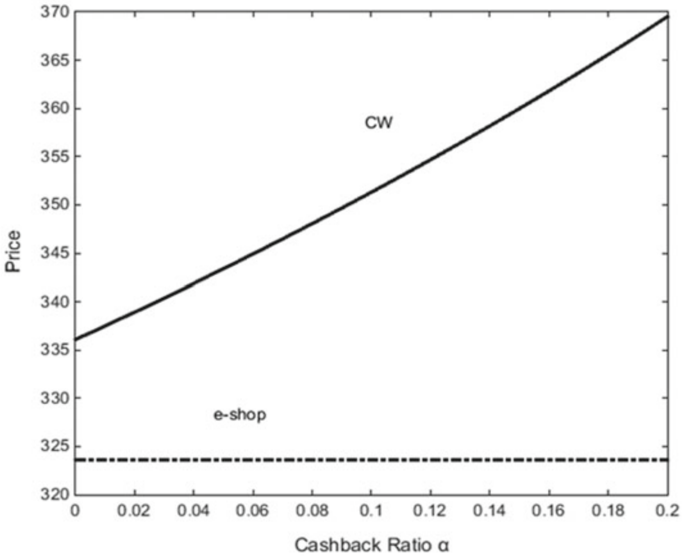


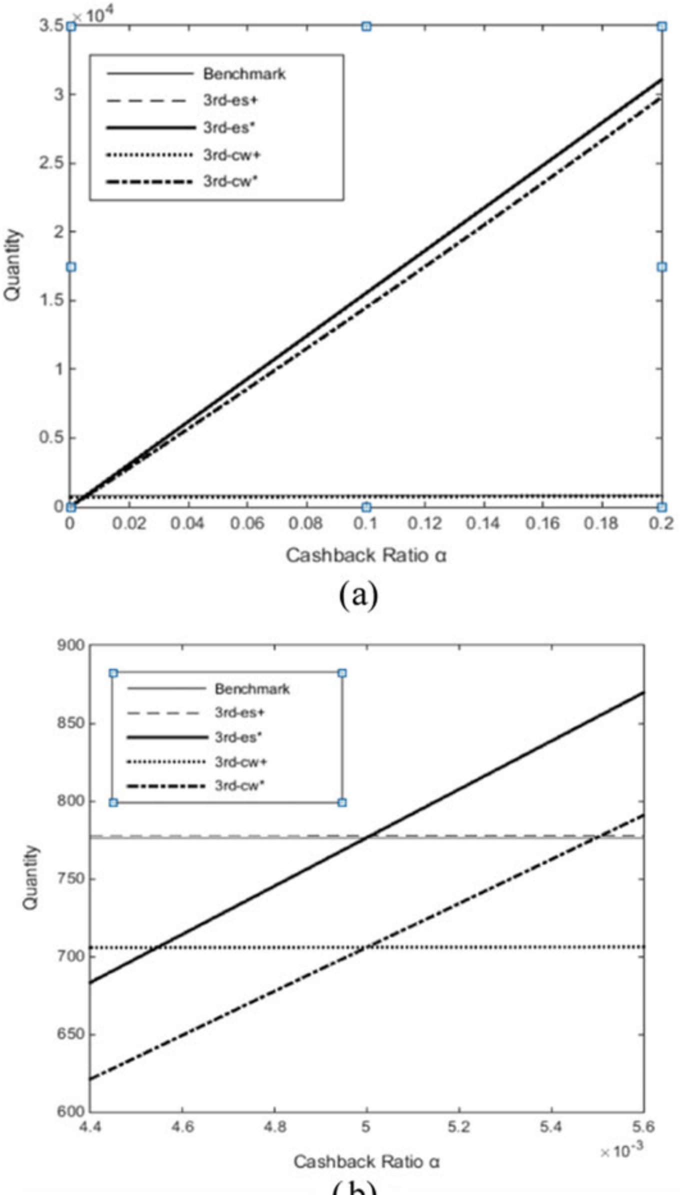
Fig. 1 The equilibrium price versus the cashback ratio α

without CW. On the whole, under multiplicative demand, the effect of cashback on order quantity is obviously higher than that under additive demand.

We can easily find in Fig. 3a that as the cashback ratio increases, the e-tailer's profit increases. Under the multiplicative demand, the e-tailer's profit increase more obviously than that under additive demand. In the case that the CW has small influence on the consumer demand (additive demand), if the CW plays the role of advertising and price discrimination, it will damage the e-tailer's profits (in the case of 3rd-cw+). However, if the CW only plays the role of advertisement (in the case of 3rd -es+), it will make the e-tailer's profit higher than the CW is not used (Benchmark). When the impact of CW on consumer demand is large (multiplicative demand), the use of CW will significantly increase the e-tailer's profit (in the case of 3rd-es* and 3rd-cw*). And e-tailer is more profitable when CW is used only for advertising (in the case of 3rd-es*).

5 Conclusion

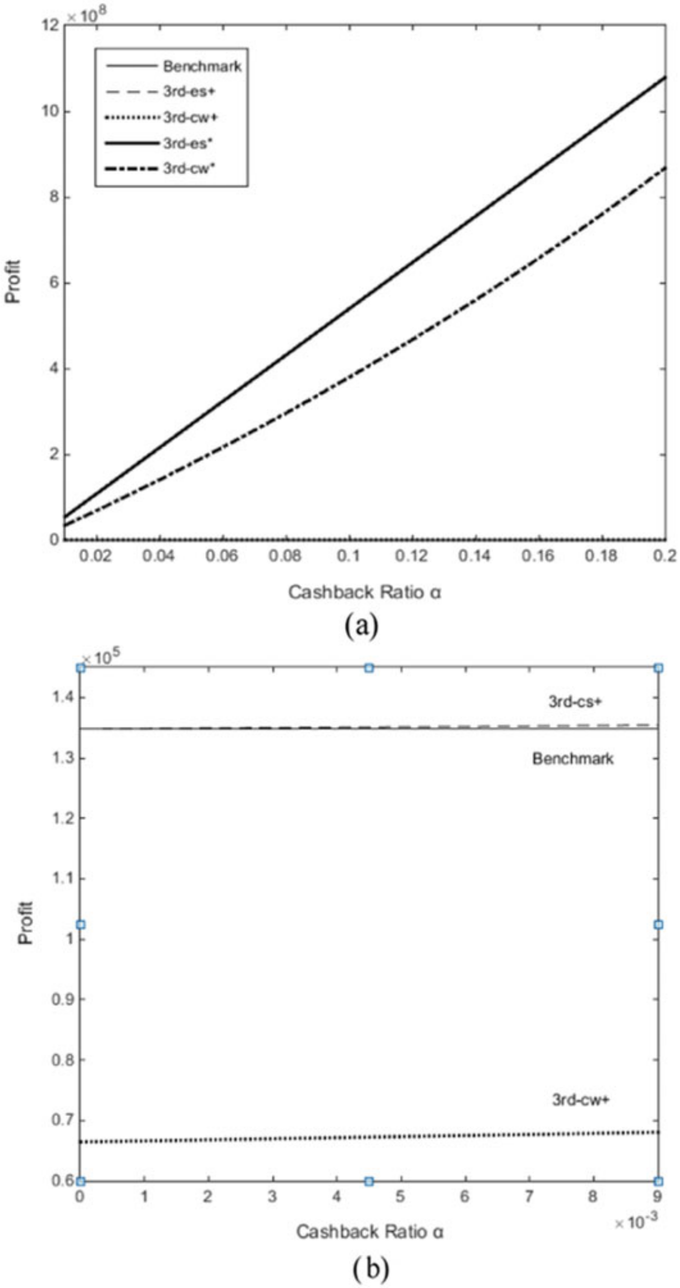
This paper extends the newsvendor model to characterize the equilibrium decisions for each participant and investigate the impacts of decentralized cashback (Third-party CW) on e-tailers' price and quantity decisions in the presence of strategic consumers. And that's where the innovation of the model comes in integrating the cashback promotion strategy and the strategic consumer's waiting behavior into a framework and considering a two-stage model. Our research leads to the following



(a)

(b)

Fig. 2 The equilibrium quantity versus the cashback ratio α



(a)

(b)

Fig. 3 The equilibrium profit versus the cashback ratio α

conclusions. First, the use of CW can stimulate demand, and restrain the waiting behavior of strategic consumers. When the CW can significantly improve consumer demand, it can significantly improve consumer profits. Moreover, e-tailers can make CW play different roles through pricing, so as to obtain better profits. Second, e-tailers can use cashback strategy to create a false image of the price cut. This means that with CW, the strategic consumer may face a higher price in full-price period than that without CW.

These findings also bring some beneficial management enlightenment for e-commerce sales First, CW can help e-tailers cope to some extent with strategic consumer waiting behavior to a degree. Second, when choosing third-party CWs for cooperation, e-tailers should choose platforms with good influence, such as public praise and popularity, so as to generate better advertising effect, stimulate demand and expand profits. Third, the setting of retail prices is crucial in determining the role played by CWs, which directly affects profits. In the presence of strategic consumers, e-tailers can make CW play different roles and guide consumers to buy through a certain channel through different pricing. Under a "pay per transaction" commission model, e-tailers can make more profits when the CW only plays the role of advertising.

In the future, we can extend our model into centralized case, that is the self-owned CW. Moreover, in reality, the payment of cashback is often delayed, the late payment deadlines would also be a point worth discussing.

Acknowledgements This work is supported by National Natural Science Foundation of China under grant 71771123.

Appendix: Proofs

Proof of Proposition 1

The first and second-order condition of (2) is as follows: $\frac{d\prod}{dQ} = (p-s)\bar{F}(Q) - (c-s)$ and $\frac{d^2\prod}{d^2Q} = -(p-s)f(Q) < 0$. Thus, we have $\bar{F}(Q) = (c-s)/(p-s)$. From (1) and (3) we obtain $p^* = s + \sqrt{(c-s)(v-s)}$, $\bar{F}(Q^*) = \sqrt{\frac{c-s}{v-s}}$. Substituting p^* and Q^* into the profit function, we get (5). ■

Proof of Proposition 2

When $v \geq \frac{\alpha p^*}{1-\theta}$, the profit function is

$$\prod_{es}(Q,p) = (p-s)E[\min\{D(\alpha,\sigma),Q\}] - (c-s)Q.$$

Taking the first-order and second-order derivatives of the profit function with respect to Q, respectively, we have

$$\frac{d\prod_{es}}{dQ} = (p-s)\bar{F}(Q-d(\alpha)) - (c-s)$$

$$\frac{d^2\prod_{es}}{d^2Q} = -(p-s)f(Q-d(\alpha))$$

Because $\frac{d^2\prod_{es}}{d^2Q} = -(p-s)f(Q-d(\alpha)) < 0$, thus, $\prod_{es}$ is concave on Q, and there exists a unique optimal solution. Therefore, we get (12).

When $v < \frac{\alpha p^*}{1-\theta}$, the profit function is

$$\prod_{cw}(Q,p) = ((1-\beta)p-s)E[\min\{D(\alpha,\sigma),Q\}] - (c-s)Q$$

Taking the first-order and second-order derivatives of the profit function with respect to Q, respectively, we have

$$\frac{d\prod_{cw}}{dQ} = ((1-\beta)p-s)\bar{F}(Q-d(\alpha)) - (c-s)$$

$$\frac{d^2\prod_{cw}}{d^2Q} = -((1-\beta)p-s)f(Q-d(\alpha))$$

Because $\frac{d^2\prod_{cw}}{d^2Q} = -((1-\beta)p-s)f(Q-d(\alpha)) < 0$, thus, $\prod_{cw}$ is concave on Q, and there exists a unique optimal solution. Therefore, we get (13) ■

The proof of *Proposition* 3 is similar to *Proposition* 2. For brevity, we omit it.

References

1. Y. Aviv, A. Pazgal, Optimal pricing of seasonal products in the presence of forward-looking consumers. Manuf. Serv. Oper. Manag. **10**, 339–359 (2008)
2. Q. Lu, S. Moorthy, Coupons versus rebates. Market. Sci., **26**, 67–82 (2007)
3. Y. Liu, D. Shaozeng, Rebate strategy to stimulate online customer reviews. Int. J. Prod. Econ. **204**, 99–107 (2018)
4. A. Arya, B. Mittendorf, Managing strategic inventories via manufacturer-to-consumer rebates. Manage. Sci. **59**, 813–818 (2013)
5. J. Zhang, The benefits of consumer rebates: a strategy for gray market deterrence. Eur. J. Oper. Res. **251**, 509–521 (2015)
6. M. Khouja, J. Zhou, The effect of delayed incentives on supply chain profits and consumer surplus. Prod. Oper. Manag. **19**, 172–197 (2010)
7. F.J. Arcelus, S. Kumar, G. Srinivasan, The effectiveness of manufacturer versus retailer rebates within a newsvendor framework. Eur. J. Oper. Res. **219**, 252–263 (2011)
8. M.T. Ballestar, P. Grau-Carles, J. Sainz, Consumer behavior on cashback websites: network strategies. J. Bus. Res. **69**, 2101–2107 (2015)
9. M.T. Ballestar, J. Sainz, J. Torrent-Sellens, Social networks on cashback websites. Psychol. Market. 33, 1039–1045 (2016)

10. P. Vana, A. Lambrecht, M. Bertini, *Cashback is Cash Forward: Delaying a Discount to Increase Future Spending*, vol. 55 (Social Science Electronic Publishing, 2018), p. 852–868
11. B. Libai, E. Biyalogorsky, E. Gerstner, Setting referral fees in affiliate marketing. J. Serv. Res. **5**, 303–315 (2003)
12. Y.C. Ho, Y.J. Ho, Y. Tan, *Online Cash-Back Shopping: Implications for Consumers and E-businesses*, vol. 28 (Social Science Electronic Publishing, 2017), p. 250–264
13. Y. Zhou, B. Cao, Q. Tang, W. Zhou, Pricing and rebate strategies for an e-shop with a cashback website. Eur. J. Oper. Res. **262**, 108–122 (2017)
14. G. Lai, L.G. Debo, K. Sycara, Buy now and match later: impact of posterior price matching on profit with strategic consumers. Manuf. Serv. Oper. Manag. **12**, 33–55 (2011)
15. X. Su, F. Zhang, Strategic customer behavior, commitment, and supply chain performance. Manage. Sci. **54**, 1759–1773 (2008)
16. M. Kabul, A.K. Parlakturk, *The Value of Commitments When Selling to Strategic Consumers: A Supply Chain Perspective*, vol. 65 (Social Science Electronic Publishing, 2019), p. 1–17

Correlation Analysis of Waterway Passenger Transport and Tourism Development in Beihai City

Xiaolei Liu, Xiaofeng Li, Wenpeng Fei, Qilong Huang, and Yin Li

Abstract Along with China to build a well-off society in an all-round way, to further improve people's requirement for better life, water transportation as a traditional way of passenger transport, Its transportation function has gradually transformed into both transportation and tourism functions. On the one hand, the acceleration of tourism rise has inspired passenger routes vitality, transform and upgrade the waterway passenger transport on the other hand also further impetus to the development of the tourism industry. This paper selects waterway transportation and tourism in Beihai city as the object, analyzes their interactive development characteristics, and establishes VAR model to empirically analyze the intrinsic correlation between them. and Judge the trend of the development of waterway passenger transport and tourism interaction, put forward the corresponding development countermeasures on this basis, to provide effective policy support for the development of waterway passenger transport in Beihai City.

Keywords Waterway passenger transport · Tourism · Correlation analysis

X. Liu (✉) · X. Li · W. Fei · Q. Huang · Y. Li
China Academy of Transportation Science, Beijing, China
e-mail: 3099700230@qq.com

X. Li
e-mail: 741718321@qq.com

W. Fei
e-mail: feiwenpengfeidan@163.com

Q. Huang
e-mail: 1046852307@qq.com

Y. Li
e-mail: 408581225@qq.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_52

1 Introduction

The existing domestic and foreign studies on the relationship between waterway passenger transport and tourism development mainly focus on two aspects. One is the integration of waterway passenger transport and tourism development, the other is the impact of transportation and tourism demand. In waterway passenger transport and tourism integration development, Zhang [1] based on typical Waterway passenger station transformation experience, such as composite function transformation and waterfront tourism function, put forward the distribution center of tourism, tourist real estate development, creative industry agglomeration, public landscape space station and port economic development community and recreational business district development path; Lu [2] analyzed the resource advantages of the development of the Yangtze river waterway tourism resources. Sun [3] proposed to establish a complete urban water system in Kaifeng city and link other tourist attractions into a system. Based on the current situation of waterway passenger transport in Zhoushan city, Liang [4] made a reasonable positioning, and put forward the development ideas of waterway passenger transport from the aspects of investment and financing system, the restructuring of passenger transport enterprises and the allocation of inter-island passenger transport capacity. Starting from the current situation of waterway passenger transport in Shandong province, Shao [5] analyzed the key problems of market operation, and put forward countermeasures from the aspects of passenger transport system, navigable waters delineation, tourism passenger terminal planning and tourism passenger transport vessels. Based on the empirical data of inland river transportation in Hangzhou City, Chen [6] constructed a model by using Eviews, SPSS and GIS software to explore the correlation between inland river transportation and hinterland transportation, space and economy. In terms of the impact of transportation and tourism demand, Bieger [7] believe that air transportation and tourism development are interrelated. Warnock-Smith [8] discussed the relationship between air passenger volume and air traffic development by taking American Caribbean tourist cities as examples. Gronau [9] discussed the reasons for the success of tourism transportation as a supply side from the perspective of tourism demand side. Taking the railway transportation industry as an example, Cong [10] established the VAR model to empirically test the interaction between the number of domestic tourists and railway passenger volume during 1988–2016. Zhang [11] studied the development of railway, highway and air transportation in China and its relationship with tourism. Huang [12], Wang [13] and Yin [14] took Xinjiang Province, Zhangjiajie City and Beijing City as tourism destinations respectively, and made an empirical study on the impact of tourism transportation on tourism industry. Wang [15], Bi [16], Zhang [17] and Li [18] took Xi'an city, Hebei Province, Yunnan province and Southwest China as examples, and used the coupling coordination degree model to empirically analyze the coupling coordination between regional transportation system and tourism economic system. Based on the impulse response function, Zhu [19] quantitatively analyzed the relationship between railway transportation and tourism economic growth.

It can be seen that there are many studies on the relationship between transportation and tourism at home and abroad. Empirical evaluation mainly focuses on the analysis of the relationship between the traffic volume and the number of tourists, while there are many analyses in highway, railway and other transportation modes, while there is basically no analysis in the aspect of waterway passenger transport [20]. In terms of the integration of waterway passenger transport and tourism, domestic and foreign studies are mainly conducted on the theoretical analysis of different geographical areas such as the whole country, a specific region, a certain province, a specific city, etc., and there is no research on the correlation between waterway passenger transport and tourism. Therefore, the actual driving effect of waterway passenger transport on tourism, whether the tourism industry pulls the economic benefit growth of waterway passenger transport in reverse, and whether there is a benign linkage effect between the two are worth in-depth discussion. This paper based on the present situation analysis, takes the waterway passenger transport and tourism as the object, selection of waterway passenger Numbers reflect the waterway passenger transport capacity, selects the number of domestic and foreign tourists to reflect the tourism demand, according to the 1998–2018 time series to establish the VAR model, the empirical test of waterway passenger transport and domestic tourism development, the interaction between the trend of the development of the study both interactive, and put forward the corresponding countermeasures and Suggestions, aims to explore the mechanism of waterway passenger transport and tourism integration development, Provide a policy tool for the development of waterway passenger transportation in Beihai City.

2 Current Situation of Integrated Development of Waterway Passenger Transport and Tourism in Beihai City

Beihai City is rich in tourism resources and has great potential for tourism development. In the past ten years, Beihai city tourism industry has made full use of its advantages, made scientific planning, integrated the tourism resources, created tourism brand, and further expanded and strengthened tourism industry in Beihai. At present, Beihai city has 10 “4A+” scenic spots and 16 “3A” scenic spots. The mainland part of the city is a peninsula shaped like a rhinoceros horn, with attractive scenery and rich tourism resources. According to the demand of creating national level demonstration zones, there is a high speed development in the integration for waterway passenger transport routes and tourism. The main Waterway passenger routes, such as Beihai to Haikou route, Beihai to WeiZhou island route, Around the island navigate routes, international cruise route, will get the characteristics of Diversified, comfortable, entertainment.

On the Beihai to Haikou route, “Beibu gulf 66” opens a new shipping mode. There are two enterprises operating the Beihai—Haikou route, among which Beihai Xinyi cruise ship Co. Ltd. invested 3 Ro-Ro passenger ships (Beihai nationality) with 1431 passenger seats and 155 parking Spaces. Hainan Strait Shipping Co. Ltd. has 2 Ro-Ro passenger ships (Hainan nationality) with 1156 seats and 107 parking Spaces. In 2018, the Beihai—Haikou route reached 139,000 passengers, maintaining a steady growth trend. In 2017, the luxurious passenger ship—”Beibu gulf 66” put into use, the tourism features get more evident, The length of “Beibu gulf 66” is 118.3 m, the height of 4 layers, carry up to 718 passengers, Speed 15 knots, 12-hour flight, The ship facilities including gym, reading room, KTV, supermarket, children’s playground, outdoor recreation, etc., in the domestic passenger liner using immersive multi-function hall, for the first time realized the organic combination of transportation and tourism entertainment.

On Beihai to WeiZhou island route, relying on the “most beautiful island” to continue to expand the scale. The route is operated by Beihai Xinyi cruise ship Co. Ltd. By the end of 2018, airline capacity get stronger, there’re 7 ships in this route including “Beiyou 12”, “ Beiyou 16”, which are the advanced high-speed passenger ship. Especially in 2018, “Beiyou 25” was put in use, which is the first aircraft for tourists sightseeing customization. As “the country’s top ten most beautiful islands”, the passenger traffic on the Beihai-Weizhou island route has shown a blowout growth. In 2018, the operating number of Beihai-Weizhou route has reached 6041 and the passenger traffic has reached 3.098 million, about twice that of 2014. According to the relevant regulations, the number of passengers arriving on the island of weizhou island in a single day is 9000 in summer/autumn, and 11,000 in spring/winter. With the development of weizhou island tourism in the future, the route will continue to grow, it is necessary to further implement the concept of high-quality development and strengthen the organic integration of shipping, tourism and ecology.

Around the island navigate routes, actively create the “Beihai sea tour” brand. With the comprehensive development of the tourism industry in Beihai city, the brand construction of “Beihai Marine sightseeing tour” has been continuously deepened. In 2011, there were 3 shipping enterprises with 539 passengers, and in 2018, there were 17 shipping enterprises with 1697 passengers. Guangxi Beibu gulf port group, Beihai yangfan pilotage cruise investment co., LTD. and Guangxi chengcihang waterway operation co., LTD. Three enterprises operate 6 round-island cruise routes, respectively: Beihai port passenger terminal—white dolphins sea area—Beihai port passenger terminal; Beihai port passenger terminal—Seaside park—Beihai port passenger terminal; Beihai port passenger terminal—three thousand sea—Beihai port passenger terminal; Beihai port passenger terminal—Gaode port bridge—Beihai port passenger terminal; Nanwan terminal-Golden bay Mangrove—Nanwan terminal; Nanwan terminal—Waisha seafood Island—Nanwan terminal.

On the international cruise route, The Beihai to Vietnam maritime passenger route was approved and opened in 1997 by the Ministry of Transport. The main route is from Beihai to Ha long bay, covering 145 nautical miles. There were several companies who operate in this route, including Beihai shipping company, Beihai Huandao tourism co., LTD., Beihai lijinda international tourism shipping company, Hong Kong Pacific (Hainan) cruise co., LTD., Heligod luxury cruise co., LTD., Beihai minghua luxury cruise service co., LTD., And Beibu gulf tourism co., LTD. Since the launch of the route, the market demand has fluctuated. In 2006 and 2007, the operation has been relatively stable, and the passenger volume has reached more than 100,000 each year. With a high reputation, the route has attracted tourists from all over the country to travel to Vietnam by ship from Beihai. However, due to the influence of the general environment and the poor operation of the enterprise, the business owners were replaced in turn since 2008, and the service quality declined, resulting in a sharp decrease in passenger sources and serious losses of the enterprise. Finally, international cruise route was suspended at the end of May 2011. In recent years, with the rapid development of China's economy, cruise tourism demand continues to expand. At present, actively promote the Beihai international cruise home port construction, according to the relative transportation and tourism industry planning, the international cruise route will be restarted and the Beibu gulf region cruise brand will be made.

3 VAR Model of the Correlation Between Waterway Passenger Transport and Tourism in Beihai City

In order to analyze the correlation between waterway passenger transport and tourism development, we selected waterway passenger volume and domestic and foreign tourist volume in Beihai city from 1998 to 2018 as variables, established the var model, and made an empirical analysis of its internal correlation.

(a) ***Trend chart***

Because time series data are selected, which may be affected by heteroscedasticity, it is necessary to logarithm the sequence. In this way, the characteristics of time series data will not be affected, and the stationary sequence can be obtained more easily. The natural logarithm of domestic and foreign passenger volume was denoted as LNDT and the natural logarithm of waterway passenger volume as LNSP, and the trend charts of two variables and two first-order difference terms were drawn.

From Fig. 1, LNDT shows a relatively obvious upward trend with the growth of years. The trend of LNSP is divided into two stages. From 1998 to 2006, the change range was not large, and in 2003, there was even a small decline. Since 2008, there has been an upward trend, especially in 2010. From the two trend charts, it can be preliminarily determined that there should be a certain correlation between the waterway transport capacity (LNSP) and tourism industry (LNDT) in Beihai City. In order to study the relationship between the two, we constructs the VAR model of Beihai city's water transport capacity and tourism industry, and makes an empirical analysis of the relationship between the two (Fig. 2).

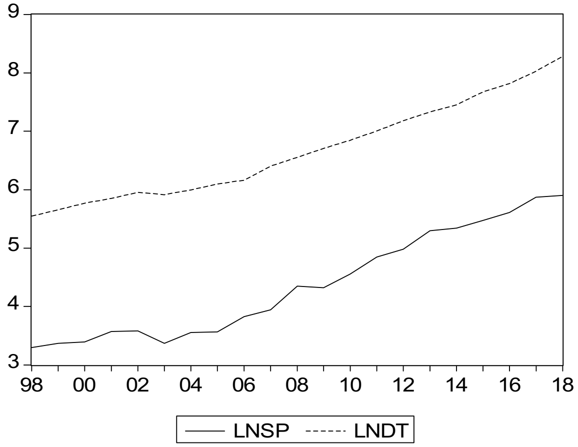


Fig. 1 Trend chart of natural logarithm of variable

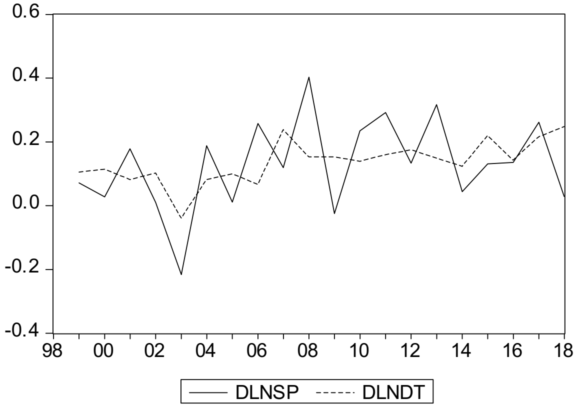


Fig. 2 Trend chart of first-order difference of Natural logarithm of variable

(b) ***stationary ADF test***
The construction of VAR model usually assumes that time series are stationary, while most economic time series are non-stationary. This was confirmed by a 1982 study by American scholars Nelson and Plosser. "Pseudo regression" may occur when linear regression is performed on non-stationary sequences. Therefore, this paper uses the unit root test (ADF) to judge the stationarity of time series. Firstly, the stationarity of LNDT and LNRP sequences was tested to avoid the phenomenon of "pseudo regression" of the sequences. The estimators obtained by ordinary least square method are guaranteed to be consistent and have an asymptotic normal distribution. See Table 1.

According to SC minimum criterion, the lag coefficient of the variable is determined, The test results show that horizontal data cannot reject the hypothesis of unit root, But the first difference of time series can reject the null hypothesis of existence

Table 1 Results of unit root test of LNDT and LNSP from 1998 to 2018

Sequential	Test type (C, T, K)	P	ADF Test value	Critical value	SC value	Single integral order
LnCE	(C, 0, 1)	0.0128	−1.023046	−2.986225	−0.706954	1
LnTRA	(C, 0, 1)	0.0294	−1.082845	−2.820619	−1.763011	1

Note C, K and T are the constant term, time trend variable and lag period in the test model. Zero is no constant term and no time trend

of unit root at the significance level of 5%. It shows that all sequences are first order stationary.

(c) ***Granger causality test***
Since the first-order difference of time series is stable, they can be tested by granger causality test to determine the causal relationship between variables. Based on the AIC minimum principle, the results of granger causality test of LNDT and LNSP were obtained, as shown in Table 2. The results show that at the granger level of 5%, the waterway transport capacity is the granger reason for the development of tourism, and the development of tourism is also the granger reason for the development of waterway transport capacity. The two are mutually causal, and the VAR model is used to further analyze the correlation between the two.

(d) ***VAR Model***
In order to establish the most reasonable and ideal VAR model, the sequence LNDT and LNSP are modeled one by one, VAR_DT_SP is denoted, The LR/FPE/AIC/SC/HQ criterion was used for the test to find the optimal lag period. The test results obtained are shown in Table 3. According to the data in Table 3, the VAR_DT_RP model is optimal when P = 2 in the lag period. On the basis of the analysis, the model VAR_DT_SP, r-squared is 0.996512, and the model has a very high degree of fit.

(e) ***Johansen test***
Johansen test was used to observe whether there is a common stochastic trend between LNDT and LNSP. The test results are shown in Table 4. It can be seen that there is a long-term stable co-integration relationship between LNDT and LNSP, and the expression of the standardized co-integration equation is shown in the formula. According to the formula, there is a long-term equilibrium relationship between the number of tourists and the amount of water transport in Beihai city: With other conditions unchanged, each percentage point increase in water transport capacity can boost the number of tourists by 1.73181 percentage

Table 2 Results of granger causality test of LNDT and LNSP from 1998 to 2018

The null hypothesis	Observations	F-statistic	Probability
LNDT is not the Granger cause of LNSP	19	2.45798	0.0187
LNSP is not the Granger cause of LNDT		1.34247	0.0504

Table 3 VAR model

	LNDT	LNSP
LNDT(−1)	0.941663	1.405352
	(0.29299)	(0.73485)
	(3.21398)	(1.91245)
LNDT(−2)	0.234071	−0.627020
	(0.27961)	(0.70129)
	(0.83714)	(−0.89410)
LNSP(−1)	0.092581	0.206374
	(0.13096)	(0.32845)
	(0.70696)	(0.62833)
LNSP(−2)	−0.197074	0.120537
	(0.12284)	(0.30809)
	(−1.60432)	(0.39124)
C	−0.571294	−2.174409
	(0.50183)	(1.25864)
	(−1.13842)	(−1.72758)
R-squared	0.996512	0.982422
Adj. R-squared	0.995515	0.977399
Sum sq. resids	0.040547	0.255061
S.E. equation	0.053816	0.134976
Log likelihood	31.46265	13.99173
Akaike AIC	−2.785542	−0.946498
Schwarz SC	−2.537005	−0.697961
Mean dependent	6.788573	4.492450
S.D. dependent	0.803629	0.897838

Table 4 Johansen inspection

Hypothesis	The eigenvalue	Trace statistics	5% critical value	1% critical value
None	0.454459	16.12355	15.41	20.04
At most 1	0.208812	4.215950	3.76	6.65

points, indicating that water transport capacity plays a certain role in promoting the development of tourism in Beihai City.

$$\mathrm{LNDT} = 45.47045 + 1.73181\mathrm{LNSP}$$

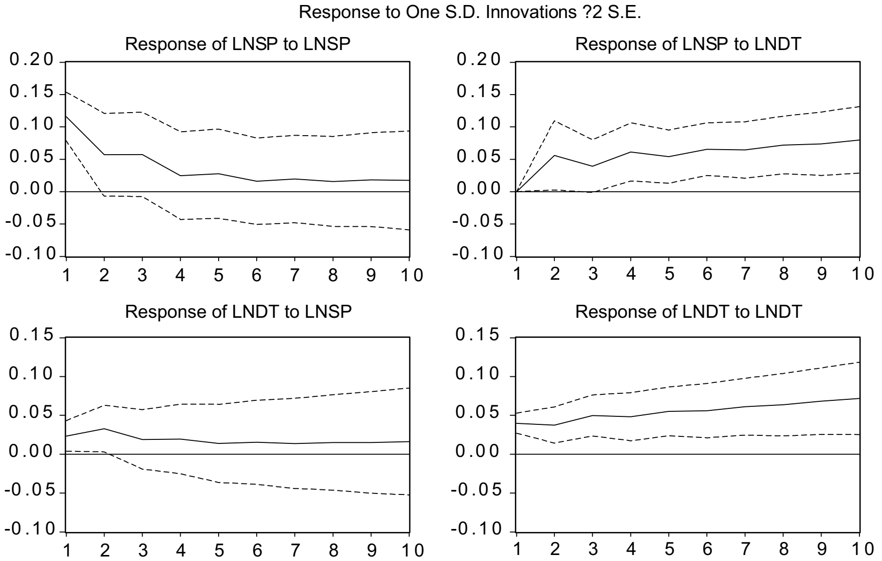


Fig. 3 Response to the impact of standard information

(f) ***impulse response analysis***
LNDT and LNSP were hit with a positive standard deviation information respectively, and the response trajectories of LNDT and LNSP under impact were observed, as shown in the Fig. 3.

Seen from Fig. 3, LNSP has no significant effect on its own impact and fluctuates greatly. In the first phase, a relatively high level of response was made, showing as 0.05; Begins to decline in Phase 2; By phase 4, it starts to go negative. The reaction of LNDT to the information of the standard deviation of LNSP is as follows: No response in phase 1; In the second phase, there was a significant increase to 0.03, followed by an upward trend; In phase 3, it dropped slightly, but continued to rise until phase 5, It reached 0.058 by phase 10. The economic implications of this conclusion are as follows: The growth of waterway transportation is also influenced by itself and domestic tourism demand, but its own impact from the beginning of a high, to the later negative impact, and the impact of tourism demand continues to rise. In other words, In the short term, waterway transportation is mainly driven by itself, but this driving effect is obviously weakened with the passage of time, while the driving effect of tourism demand in Beihai city is continuously strengthened. In the long run, the driving force of tourism demand will exceed its own influence and become the leading force to promote the development of the water transport industry.

The disturbance of LNDT on its own standard deviation showed a positive effect, which was reflected as 0.039 at the beginning of phase 1, and then presented an increasing change pattern, reaching 0.104 in phase 10. The impact of LNSP on LNDT did not respond in the phase 1, and the growth rate began to slow down in

Table 5 Variance decomposition of LNDT

Phase	The standard deviation	LNDT	LNSP
1	0.046196	100.0000	0.000000
2	0.067941	98.15196	1.848045
3	0.086300	97.75315	2.246845
4	0.100777	97.76607	2.233925
5	0.115642	96.37072	3.629285
6	0.129400	95.70030	4.299702
7	0.143733	94.71520	5.284796
8	0.157915	94.12280	5.877205
9	0.172669	93.50219	6.497813
10	0.187707	93.07390	6.926102

the phase 2, which was significantly weaker than the impact of tourism disturbance in the phase 10. The economic implications of this conclusion are as follows: The number of tourists in Beihai city is jointly promoted by itself and the waterway transport industry, but it has a greater influence on it. In other words, the demand of the tourism industry in Beihai city is more dependent on its own influence.

(g) ***variance decomposition***
In order to further understand the interaction between tourism and waterway passenger transport, VAR_DT_SP model was decomposed by variance, and the contribution of each structural impact to the change of LNDT and LNSP was analyzed, To evaluate the importance of the interaction between the two, the output results are shown in Tables 5 and 6.

According to the data in Table 5, LNDT was mainly influenced by itself. It was 100% at the beginning of phase 1, and continued to decline, but remained at 93%

Table 6 Variance decomposition of LNSP

Phase	The standard deviation	LNDT	LNSP
1	0.046196	25.86129	74.13871
2	0.067941	47.56619	52.43381
3	0.086300	54.33079	45.66921
4	0.100777	60.86021	39.13979
5	0.115642	65.22412	34.77588
6	0.129400	68.43556	31.56444
7	0.143733	71.23713	28.76287
8	0.157915	73.39843	26.60157
9	0.172669	75.40271	24.59729
10	0.187707	77.05058	22.94942

in phase 10. The influence of LNSP on its prediction variance increased from the second phase, but only reached 22.9% by the phase 10. It shows that over the long term, 22.9% of the tourism demand in the north sea is explained by the amount of water transport, which proves that the amount of water transport plays a small role in promoting the tourism demand in the north sea and there is still much room for improvement. From the data in Table 6, it can be concluded that LNSP is affected by the dual impact of LNDT and its own fluctuations at the beginning, but it is at a high level (74.139%). The influence of LNDT on it showed a strong upward trend from phase 1, and reached 77.05% in phase 10, which exceeded the influence of LNSP on it. It can be seen that the tourism demand in Beihai city has a strong stimulating effect on Waterway passenger.

(h) ***empirical conclusion***

(1) The waterway transport capacity is the granger reason for the development of tourism, and the development of tourism is also the granger reason for promoting the waterway transport capacity. They cause and effect each other.
(2) The waterway transport capacity is a long-term co-integration relationship between the number of tourists in beihai city, and the waterway transport capacity plays a certain role in promoting the latter.
(3) In the long run, the tourism demand of Beihai city continuously improves the demand of waterway transport capacity, and the stimulation effect is strong. The stimulation effect of tourism is higher than that of waterway transport itself. However, the waterway transport capacity does not play a significant role in promoting the tourism demand of Beihai city, so there is still a large room for improvement. Tourism demand is more dependent on its own role, the waterway transport industry did not play its due role of linkage promotion.

4 Trend Analysis of Waterway Passenger Transport and Tourism Integration in Beihai City

From 2010 to 2018, the number of tourists in Beihai City grow rapidly, among which the number of inbound tourists increased from 73,000 to 160,000, nearly tripling, with an annual growth rate of 10.36%. The number of domestic tourists has also increased significantly on the basis of 9.38 million. In terms of the growth trend of tourists, the growth rate of domestic tourists is higher than that of inbound tourists, accounting for more than 99% of the total number of tourists. It can be seen that Beihai City has gradually become a tourist resort for domestic tourists in recent years. The increase of tourists will inevitably drive the growth of the urban tourism income, since 2010, the passengers tourist foreign exchange income has increased by 3.35 times, more than $70 million, domestic tourism income is increased by more than 7.46 times, reached 49.967 billion yuan, the tourism industry has been a rapid growth spot of economy. With the proposal of the national "One Belt And One Road" initiative, as one of the ports on China's ancient maritime silk road, Beihai City is an important node city connecting Asean countries. Tourism is the shining name card of Beihai City, ten miles of silver beach, one hundred years old street, one thousand acres of mangrove, ten thousand years of island. To promote the development of tourism industry, Beihai city introduced a series of supporting policies such as construction of international tourism hotels, vigorously promote the tourism project, comprehensively improve the tourism environment. Beihai will be built into a core city of Beibu gulf international tourism resort and a domestic first-class and internationally renowned coastal tourism and leisure resort.

In the future, with the deepening of the strategies of China-Asean free trade area, Guangdong-Hong kong-Macao greater bay area, Hainan free trade port, and new routes in the west china, the number of foreign tourists in Beihai will be further increased. Under the stimulation of tourism, the waterway transport industry in Beihai city will usher in a new round of development, in Beihai-Haikou route, Beihai-weizhou island route, Around the island navigate routes, international cruise route, showing a trend of steady development.

From the perspective of Beihai to Haikou route, since 2010, the annual passenger volume of the route has increased from 175,000 to 278,000, with a growth rate of 58.9% and an average annual growth rate of 6.0%. The passenger volume between the two sides is relatively balanced, which further indicates that the passengers from the two sides are common. From the perspective of age structure, with the investment of new ships in 2017, the aging status of ships has been improved. In the future, the route will be further invested in high-speed ro-ro passenger ships, and the time from Beihai to Haikou will be shortened to 5.5 h, so as to meet the travel needs of tourists from southwest China to Hainan province. At the same time, with the further balance of passenger flow and the optimization of transport capacity, the tourism attraction of Beihai city will be further enhanced, and stimulate the growth of the waterway passenger transport market. From the perspective of future development, Beihai should strengthen the integrated development of waterway passenger transport with Yangpu, Sanya and other cities in Hainan province, give play to the agglomeration effect of tourism factors, and deepen the integrated development of tourism and transportation.

From the perspective of Beihai to WeiZhou island route, since 2010, the shipping capacity of the route has increased from 4 vessels to 7 vessels, with an increase rate of 75%. Passenger traffic increased from 694,000 to 3.098 million, a growth rate of more than 346% and an average annual growth rate of 20.6%. The annual number of vessels also increased from 3980 to 6041, an increase of 51.8%. Based on the overall analysis of the passenger volume of Beihai waterway, the passenger volume of this route accounts for more than 80% of the passenger volume of Beihai waterway. From the analysis of route positioning, tourism as the leading; From the perspective of ship choice, passengers can choose to take high-speed passenger ship or ordinary passenger ship. In the future, with the further development of weizhou island and its surrounding tourism resources, the passenger volume of this route will continue to grow. The number of passengers on the island in a single day is limited to 11,000, and the total passenger transport is limited by the carrying capacity of environmental resources.

From the perspective of Around the island navigate routes, since 2014, the routes have shown an overall growth trend, with 366,600 passengers in 2018, an increase of 43.49% and an average annual growth rate of 4.99% compared with 251,300 in 2014. In the future, the silver beach, home of oversea Chinese from Vietnam, the seaside park, one hundred old street and other traditional projects will continue to keep the heat, and three thousand sea, white dolphins scene, Hester's first port, Quzhang six lake and star islands lake, will Create new tourist hotspots. With the construction of ports such as Hepu port of departure, Waisha port, Xicun port, Yingpan port and Shatin port, Beihai city will continue to enrich the passenger routes to further meet the different needs of tourists and the passenger volume will continue to grow.

From the point of international cruise route, with the deepening of the "area" initiative, as well as effects appeared from Guangdong-hong kong-macao greater bay area, Hainan free trade port, New land and sea lanes in the west. The tourism resources development of Hainan province and southeast Asian nations has become the inevitable direction, provides conditions for the recovery of international cruise route from Beihai. At the same time, Sansha route, as the blue sea of China's cruise route, also become an important factor to promote the development of cruise route in the Beibu gulf region. Formally began in 2017, Beihai cruise home port has officially started construction. This project is the core project to implement the transformation and upgrading of beihai port. It has been included in the 2017 national list of selected tourism projects of the ministry of culture and tourism, and is a key project of 10 billion yuan at the level of the autonomous region. In the future, it will integrate the cruise line resources of Beihai, Sanya, Sansha and southeast Asia, and gradually promote the opening of cruise route to call at Guangdong-hong kong-macao greater bay area and Hainan free trade port, so as to create high-end cruise tourism products in Beihai and further enrich the aquatic tourism products in Beihai.

5 Suggestions on the Interactive Development of Waterway Transportation and Tourism in Beihai City

(a) ***accelerate the adjustment and optimization of route capacity structure***
Actively eliminate old ship types, invest in high-speed passenger ships, shorten the route running time, to promote the rapid, large-scale, comfortable and recreational development of Beihai-Haikou route, upgrade the Beihai to WeiZhou island route to high-speed passenger ships and advanced sightseeing cruise ships., Upgrade the capacity structure of round-island tour, Beihai water sightseeing tour has become a representative product of beihai tourism. On the basis of Vietnam's Halong bay international cruise, introduction of luxury cruise

ship, dock in Guangzhou, Shenzhen, Hong Kong and Sanya cruise home port, in order to "China-vietnam trade and cultural exchanges", "Beautiful Sansha" as the theme, set up the silk road theme cultural park and Vietnam sea trade market stagnation, reconstruct the "sea silk road".

(b) ***establish a coordinated and unified passenger route service network.***
According to the construction of xi "human destiny community" concept as the guidance, take the people first, sets up the concept of collaborative development, Actively coordinate with provinces and cities along the Qiongzhou strait such as Hainan and Guangdong, In line with the principle of "consultation, coordination and win-win", On the construction of market mechanism, such as passenger flow, shipping capacity and flow direction, to realize coordinated operation and promote the sound development of the Beihai to Haikou route.

To form the island tour management committee, carry out the "4 unified" management of the route around the island, that is, unified planning and construction, unified scheduling and management, unified marketing, unified window sales. Reasonable arrangement of travel routes, improve the service quality of the route, ensure the economic benefits of enterprises, make the tourism market in order; Through the organization of marketing team, expand the publicity, open up channels of market customers. At the same time, we will push forward the reorganization and merger of four enterprises that travel around the island, and integrate them into one or two joint ventures. We will adjust and control the total capacity, optimize the capacity structure, to meet the requirements of the development of the island tour market.

In the process of construction of international cruise route, focus on connecting Sanya, Guangzhou, Xiamen cruise home port, And actively dock with the cruise home ports of Vietnam, Thailand, Malaysia and other southeast Asian countries. In the infrastructure supporting, cruise line market operation, product promotion and other aspects of coordination and unity, vigorously develop the "one journey, multiple stations" cruise products.

(c) ***accelerating the construction of high-speed passenger ship terminals***
Through the consultation and communication between Guangxi district and Beihai city and the Hainan provincial government, Aiming at the problem that Hainan province does not yet have the berthing port for high-speed passenger ships of the Beihai-Haikou route, To implement the planning, site selection and construction of the berthing dock for high-speed passenger ships in Hainan province as soon as possible, so as to realize the navigation of high-speed passenger ships at an early date. At the same time, aiming at the construction of the collection and distribution channel of Yangpu port passenger terminal, actively promotes the construction of the collection and distribution system of Yangpu port, so as to realize the seamless connection between waterway transportation and highway transportation.

According to the international standards, accelerate the renovation and expansion of the international passenger port and the passenger station at the west corner of Weizhou island, integrate tourism culture, tourism catering, tourist attractions and tourism transportation into the passenger station construction, and actively draw on the experience of famous scenic spots such as Hainan wuzhizhou island, Xiamen gulangyu island and Guangzhou Chime-long in the operation and organization mode. We will improve the construction level of the overseas Chinese international passenger port and the west corner passenger terminal, improve the waterway and berthing conditions, and meet the requirements of the new generation of large and advanced sightseeing cruise ships.

Drawing on the experience of water tourism transportation in internationally developed cities such as Tokyo and Hong Kong, Cooperate with the key projects of hepu county shankou wetland ecological tourism area, white dragon comprehensive leisure tourism area of Tieshan port, Hepu historical and cultural tourism area, etc. To speed up the layout of the Around the island navigate routes, By connecting the key tourist areas with the routes around the island, the attraction of tourism products to the passenger flow can be maximized.

(d) ***vigorously promote energy-saving and environment-friendly ships***
The Beihai to Weizhou island route should be priority develop environment-friendly ships with less energy consumption and low pollution, to minimize pollutant discharge and protect natural ecological resources. In accordance with the relevant support policies of the state, we will grant financial subsidies and tax incentives.

The sailing conditions of the around island route are better, and requirements for passenger ships are relatively low. Among the small and medium-sized passenger ships, the energy-saving and environment-friendly ships can be actively promoted to reduce energy consumption and pollutant emission, so as to effectively maintain the coastal ecological environment of Beihai City.

6 Conclusion

Based on the dynamic analysis of waterway passenger transport and tourism development in Beihai city, this paper studies the internal correlation between waterway transport and tourism development and draws the following conclusions:

(1) From the current situation, the development of waterway passenger transport in beihai city shows the characteristics of tourism development;
(2) In terms of correlation, waterway passenger transport and tourism development in beihai city are granger causality. Among them, the development of tourism in beihai city stimulates the capacity of waterway passenger transport. However, the stimulation effect of water transport on the development of tourism in Beihai city is not obvious;
(3) From the trend analysis, waterway passenger transport in beihai city will show a significant growth in the future under the promotion of tourism. Beihai-Haikou

route, Beihai-weizhou island route, Around the island navigate routes, international cruise route, Will become the main force to support the development of waterway passenger transport and tourism;

(4) It is suggested to accelerate the adjustment and optimization of route capacity structure, establish a coordinated and unified passenger route service network, accelerate the construction of high-speed passenger ship terminals, and vigorously promote energy-saving and environment-friendly ships. Promote the development of waterway passenger transport in beihai city, and then effectively support the development of tourism.

Through field research and empirical analysis, this paper studied the internal relationship between waterway passenger transport and tourism development in Beihai city, and gave corresponding development Suggestions. Limited by the time of investigation and the availability of data, and limited variables to select. This paper aims to establish a discussion basis for the empirical analysis of the interaction between waterway passenger transport and tourism.

References

1. J. Zhang, Y. Yang, Path and key issues of tourism function transition for waterway passenger terminal. J. Shanghai Marit. Univ., p. 7–13 (2011)
2. L. Lu, Shallows Peaker river water transportation traveling development resources superiority. China Water Transp. (Academy Version), p. 153–155 (2006)
3. J. Sun, S. Yang, Preliminary research on the development of Kaifeng urban tourism water system resources system. J. Yellow River Conserv. Tech. Inst., p. 62–65 (2001)
4. S. Liang, Countermeasures to promote the development of waterway passenger transport in Zhoushan island. China Water Transp., p. 47–49 (2008)
5. Q. Chen, Y. Dong, J. Zhou, Correlation between waterway and inland in Hangzhou. J. Zhejiang Univ. Technol. (Social Science), p. 38–44 (2016)
6. M. Shao, Analysis on the problems and countermeasures of waterway tourism passenger transport in Shandong Province. China Port, p. 51–53 (2019)
7. T. Bieger, A. Wittmer, "Air Transport and Tourism: Perspectives and Challenges for Destinations, Airlines and Governments. J. Air Transp. Manag. **12**(1), 40–46 (2006)
8. D. Warnock-Smith, P. Morrel, Air transport liberalisation and traffic growth in tourism-dependent economies: a case-history of some US-Caribbean Markets. J. Air Transp. Manag. **14**, 82–91 (2008)
9. W. Gronauw, A. Kagermeier, Key Factors for Successful Leisure and Tourism Public Transport Provision. J. Transp. Geogr. **15**(2), 127–135 (2007)
10. C. Cuili, Research on the interaction between domestic tourism and railway transportation industry in China based on VAR model. J. Hunan Univ. Technol., p. 66–71 (2018)
11. J. Zhang, L. Lu, Wuhu Changjiang Bridge and the improvement of tourist traffic condition in Anhui Province. Hum. Geogr. **17**(4), 75–79 (2002)
12. L. Huang, *Xinjiang Highway Traffic on the Development of Tourism* (Xinjiang Normal University, 2008)
13. Z. Wang, The research of the influence of the tourism transportation to the tourism development in Zhangjiajie. Theory Pract. Financ. Econ. **30**(4), 112–116 (2009)
14. P. Yin, Study of the influence of tourism transport cost on tourism destination spatial competition. Areal Res. Dev., **31**(6), PP. 87–91 (2012)

15. Y. Wang, Y. Ma, Analysis of coupling coordination between urban tourism economy and transport system development: a case study of Xi'an City. J. Shaanxi Normal Univ. (Natural Science Edition) **39**(1), 86–90 (2011)
16. L. Bi, Y. Ma, Analysis of coupling coordination between traffic system development and Province tourism economy: a case study of Yunnan Province. J. Xi'an Univ. Financ. Econ. **26**(1), 124–128 (2013)
17. S. Zhang, F. Wei, Study on the interaction between tourism economy and traffic optimization based on the coupling degree model: a case study of Hebei Province. Shanxi Agric. Sci. **58**(4), 163–166 (2012)
18. X. Li, *The Research of Coordinated Development of Tourist Flow and Transport in Southwest China* (Jishou University, 2013)
19. T. Zhu, J. Lu, Z. Zhu, Study on relationship between railway traffic and tour economic growth in China based on impulse response function. Railw. Transp. Econ. **37**(7), . 54–60 (2015)
20. A. Erdil, An overview of sustainability of transportation systems: a quality oriented approach. Tehnicki vjesnik-Technical Gazette **25**(2), 343–353 (2019)

Towards Fully-Synthetic Training for Industrial Applications

Christopher Mayershofer, Tao Ge, and Johannes Fottner

Abstract This paper proposes a scalable approach for synthetic image generation of industrial objects leveraging Blender for image rendering. In addition to common components in synthetic image generation research, three novel features are presented: First, we model relations between target objects and randomly apply those during scene generation (Object Relation Modelling (ORM)). Second, we extend the idea of distractors and create Object-alike Distractors (OAD), resembling the textural appearance (i.e. material and size) of target objects. And third, we propose a Mixed-lighting Illumination (MLI), combining global and local light sources to automatically create a diverse illumination of the scene. In addition to the image generation approach we create an industry-centered dataset for evaluation purposes. Experiments show, that our approach enables fully synthetic training of object detectors for industrial use-cases. Moreover, an ablation study provides evidence on the performance boost in object detection when using our novel features.

Keywords Object detection · Synthetic data · Domain randomization

1 Introduction

In recent years, machine learning methods have gained increasing attention. Particularly, supervised learning using deep neural networks solves previously insoluble problems. These developments are especially apparent in the field of computer vision: convolutional neural networks (CNNs) enable object detection [1–4] and segmenta-

This work was supported by the BMW AG (BMW Group).

C. Mayershofer (✉) · T. Ge · J. Fottner
Chair of Materials Handling, Material Flow, Logistics,
Technical University of Munich, Garching, Germany
e-mail: christopher.mayershofer@tum.de

T. Ge
e-mail: tao.ge@tum.de

J. Fottner
e-mail: j.fottner@tum.de

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_53

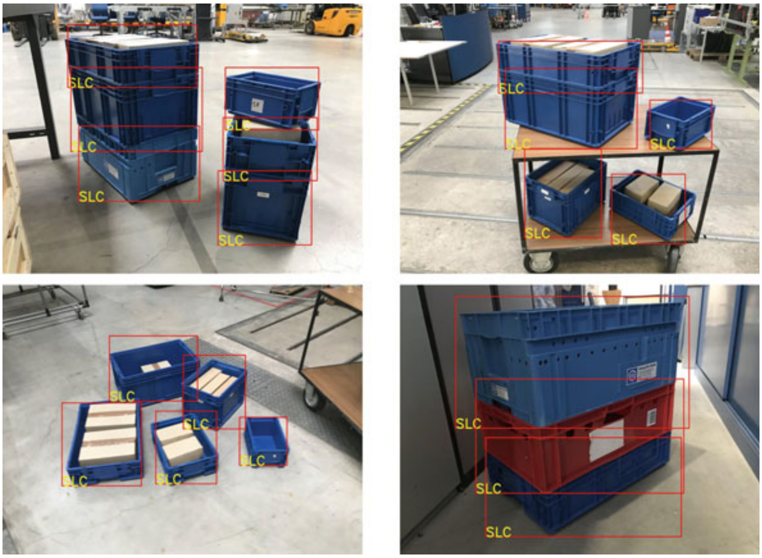


Fig. 1 Prediction results of a fully-synthetic trained object detector. The detection model is trained on *synthetic image data only* using the proposed image generation approach. To meet industrial requirements, we implement novel features to represent specific object relations, to suppress false detections and to model complex industrial lighting conditions while maintaining a maximum level of scalability

tion [5, 6], as well as pose [7, 8] and depth estimation [9, 10]. A decisive factor for the success of these networks is the existence of large amounts of annotated image data.

While the global research community has already published many diverse datasets, there are still use cases for which no or insufficient data is available. In the public sector in particular, there are large datasets available that deal with autonomous driving [11, 12], common objects [13, 14], or famous landmarks [15], to only name a few. On the other hand, datasets within the industrial domain are rarely found, as few images are published. Accordingly, for many industrial applications, realistic datasets must first be collected and annotated before applying them to specific use-cases. Collecting and annotating a dataset is a very money and time consuming endeavour [16].

The time required for annotating images depends on the specific application. For an image-level classification task, for example, annotation can be done rather quickly, as it means that each image to be used for training needs to be assigned to a specific class. The more complex the task, the greater the annotation effort; for example, within object detection, in addition to the classification of an object, its position in the image plays a decisive role. Depending on the desired level of detail, the localization can be done using bounding boxes (low level of detail resulting in low annotation effort) or segmentation masks (high level of detail resulting in high annotation effort).

To solve this challenge in a long-lasting manner, different methods were proposed to synthetically create image data in order to train deep neural networks. In contrast to generating training datasets using natural images, artificial images can be created automatically with very precise annotations (i.e. bounding box, per-pixel depth, object pose, object segmentation, etc.) at negligible cost. However, the existing domain gap between synthetic and natural images makes the neural network trained on synthetic images perform poorly on natural images [17]. Various methods have been introduced in an attempt to close this gap. Yet, there is no universal approach known to resolve this issue. Therefore, within the scope of this paper we have investigated synthetic image generation for industrial applications.

In particular we present a scalable, Blender-based image generation approach, that enables fully-synthetic training of object detectors used in industrial applications (see Fig. 1).

The scientific contributions of this paper can be summarized as follows:

(1) **Scalable synthetic image generation approach**. We present a scalable image generation approach adopting well-working methods and augmenting them with novel features such as Object-alike Distractors (OADs), Object Relation Modelling (ORM) and a Mixed-lighting Illumination (MLI).
(2) **Industry-centered evaluation dataset**. We present a natural image dataset enabling the evaluation of synthetic image generation approaches for industrial objects. Our dataset contains realistic images covering a single industrial object (small load carrier) in a close-to-industry environment facing multiple industry-relevant challenges (e.g. different lighting conditions, multiple objects, object relations).
(3) **Extensive ablation study**. We evaluate our approach on the previously mentioned dataset providing realistic performance feedback. Furthermore, we provide insights into the different aspects of our approach by ablating features and showcasing their performance boost.

2 Related Work

Generating artificial training data is gaining popularity in computer vision research. Reference [18] created training images by cutting images of target objects from other datasets and pasting them on background images. Although their method ensures patch-level realism, it requires plenty of real images with considerable human efforts in segmenting out objects. Others are using computer graphics render engines to generate synthetic training images, relying on an elaborately, manually-created and close-to-real-world scene [19, 20]. Even though synthetically generated images can appear photo-real to us humans, deep neural networks still show problems when being transferred from the simulation to the real world. The domain gap between the simulated training images and the real world experiments might be due to the fact that most render engines are built to leverage the human perception system in order to efficiently create images that appear to be photo-real [21].

Domain adaptation (DA). One way to overcome this gap is to adapt one area to another. The basic idea of DA is to transfer a model trained in the source domain to the target domain [22]. Among all different DA methods, semi-synthetic training is a simple and effective method [23]. Training the neural network on a large synthetic dataset first and fine-tuning it on limited data from the target domain afterwards boosts the performance of neural networks [16, 24, 25]. Furthermore, generative adversarial networks can be applied to achieve domain adaptation [26]. Differently, [27] focused on image translation of synthetic images. They created a generator to translate both synthetic and real images to canonical images. The application only needed to handle canonical images and never got in contact with 'raw' synthetic images. Although DA can improve neural network performance, the deficiencies are apparent. Firstly, it is inevitable to use data from the target domain. However, one of the main reasons to generate synthetic training data in the first place is to avoid the usage (and need) of real training data. Secondly, domain adaptation may improve the performance of the neural network on images from the target domain. However, when it is tested on other domains (even the source domain), its performance degrades significantly [25].

Domain randomization (DR). A second approach towards overcoming the sim2real gap is by applying DR. The general idea behind DR is that by randomizing parameters in the source domain (i.e. simulation), the target domain (i.e. real world) appears to the neural network as just another variation of the source domain [21, 28, 29]. In practice, different parameters of objects, background, camera and lights are randomized in the synthetic image generation process [24, 30]. Besides, [29, 31] imported random distractors to create random occlusion and prevent the neural network from detecting distractors in the real world. Reference [20] proposed structured domain randomization which takes the context into consideration resulting in high recall performance. Moving from color images to depth images, [28] proposes synthetic depth data randomization to generate depth images for training. Although their neural network trained only on synthetic depth images outperforms the detector trained on real data, the method is limited to depth images. Instead of generating images using modeled scenes or images as background, [21] randomly fills the background with plenty of objects in order to prevent the neural network from learning a certain pattern in the background. Their experiments demonstrate the effect of a randomized background generation.

Guided domain randomization (GDR). As an improvement to general DR, many guided DR methods were proposed. Reference [32] developed active domain randomization, which searches for the most informative environment variations by measuring the discrepancies between the randomized and reference environments. Environments with high informativeness can then increase the difficulties of training in order to improve the performance. Although active domain randomization achieves better results compared to general DR methods, the correlation between discrepancy of environments and training difficulty is still unknown. Similarly, [33] proposed the automatic domain randomization approach to increase the difficulty during the training process. This is achieved by automatically and gradually expanding the distribution over environments, helping to improve prediction accuracy but also significantly increasing the process duration.

In summary, different approaches for generating synthetic data are known. Each approach proposes novel features, which in turn are being utilized in the next iteration of synthetic image generation approaches. Since the industrial application of synthetic image data has hardly been researched so far, this paper presents our approach towards synthetic image generation for industrial applications.

3 Synthetic Image Generation of Industrial Objects

We propose a scalable image generation approach for industrial objects using computer graphics. Specifically, we render images from automatically generated 3d scenes. Hereby, we adopt well-working methods and augment them with novel features such as *Object-alike Distractors (OADs)*, *Object Relation Modelling (ORM)* and a *Mixed-lighting Illumination (MLI)*. Rendering is based on the open-source 3D creation suite Blender. Our image generation approach is visualized in Fig. 2 and can be divided into three process steps, namely background creation, foreground creation and rendering. All process steps are subject to domain randomization in order to reduce the resulting domain gap. Building on top of 3D modeled objects, we are able to automatically generate annotated training data for different computer vision tasks with varying complexity. The following section describes each process step in detail.

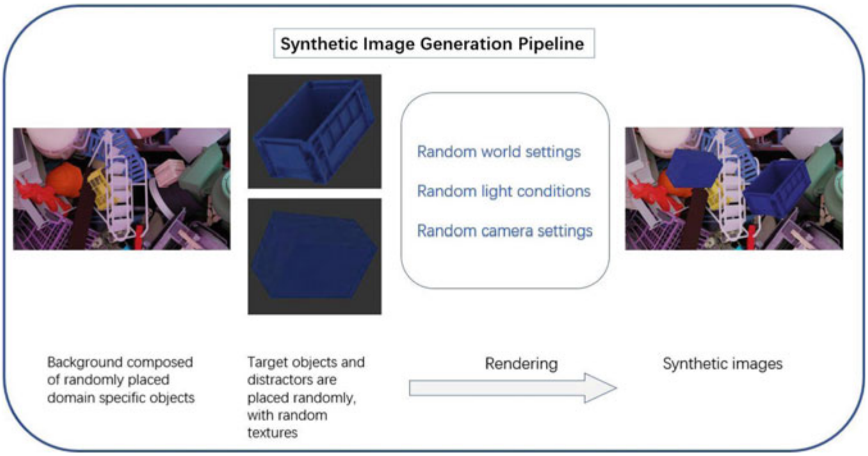

Synthetic Image Generation Pipeline
Random world settings
Random light conditions
Random camera settings
Background composed of randomly placed domain specific objects
Target objects and distractors are placed randomly, with random textures
Rendering
Synthetic images

Fig. 2 Synthetic image generation approach for industrial objects. Using 3D modeled objects, we automatically generate a 3D scene and render images from it. Wherever possible we are using domain randomization to decrease the domain gap

3.1 Background Generation

First, the background of the 3D scene is created automatically. We mostly adopt the background generation process described in [21]. In a nutshell, [21] is forming the background by using a multitude of 3D objects and randomly positioning them in the background plane.

As we found that using many 3D objects for background generation increases the computational effort and reducing the amount of 3D objects resulted in spots without any object, we additionally load and place a random image in the background plane. The voids among these objects are then filled by the loaded image. This simple measure provides a balance between the level of clutter and the computational cost for image generation.

3.2 Foreground Generation

Second, we automatically create the foreground consisting of randomized target object(s) and distractor(s) within the cameras view space. Figure 3 visualizes an automatically generated, pyramid-like 3D scene in Blender.

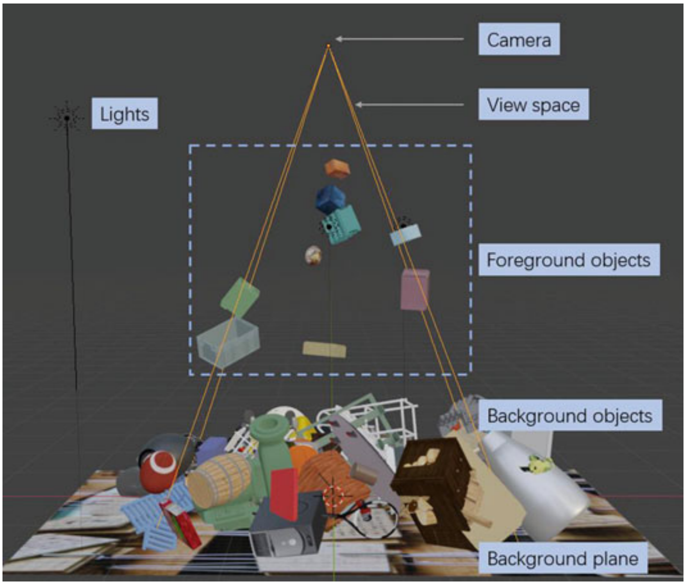


Fig. 3 Automatically generated 3D scene. Our approach creates the background using randomly placed background objects and a background plane, places target objects and distractors within the pyramid-like camera view space and creates the Mixed-lighting Illumination. Next, the 3D scene is transformed to the image space using one of Blenders' render engines

Similar to known approaches, we randomize the number, type, location, rotation and material of target objects. In addition to that, we propose *Object Relation Modelling (ORM)*, a novel feature applying predefined relations between target objects when placing them in 3D space. In order to do so, relation files are created manually by recording the relative translation and rotation of one target object to another before the image generation. This spatial relation can later be randomly applied during foreground generation. If an intersection is detected when applying a relation, the related object will be deleted. Since ORM is based on coordinate transformation, it is applicable to describe any relation amongst objects. ORM intentionally increases the probability of certain spatial relations which could hardly be achieved by chance within the synthetic dataset. As expected, our experiments show that this increases detection performance when facing those relations in natural images.

In the fashion of [29, 31], we place distractors in our scenes. Distractors are random geometries in the forground creating occlusions and 'distracting' the detector to be trained. The distractors are automatically created and placed in the scene with random geometry, position, orientation, scale and material. Additionally, we extend the idea of general distractors and implement so called *Object-alike Distractors (OADs)*. Again, OADs are randomly generated basic geometries, but in contrast to standard distractors they share the same size and material as the target object, causing the detector to focus on structural features of the target objects geometry rather than much simpler textural cues of a certain material.

3.3 Lighting and Camera

In addition to the previously mentioned steps, lighting as well as camera settings are subject to DR. The lighting condition within training data significantly affects the performance of neural networks [21, 31, 34]. In contrast to other approaches that create a single domain-randomized light source we propose a *Mixed-lighting Illumination (MLI)* that divides illumination into a global and a local component. Global lighting illuminates the entire scene whereas local lighting creates specific highlights on random positions within the cameras view space.

Global illumination consists of on the one hand passive illumination from the set environment texture and on the other hand a global light source that illuminates the entire scene. This global light source is placed randomly on a projected hemisphere. In contrast, local light sources are only placed within the pyramid-like camera view space. The environment texture as well as the global and local light source properties (i.e. location, color, energy, size) are all subject to domain randomization. Figure 4 showcases the variance in illumination; (a) shows a brightly illuminated scene with a soft greenish-turquoise hue, whereas (b) is a much darker lit scene, that puts the focus on the yellow sphere-shaped geometry due to local spotlights automatically generated by MLI.

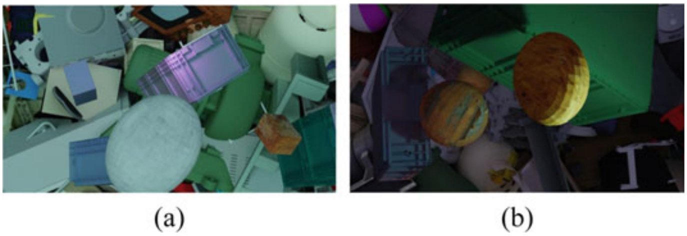


Fig. 4 Mixed-lighting illumination. Two automatically generated sample images using the mixed-lighting illumination approach. Image **a** consists of a bright scene with greenish-turquise hue, whereas **b** is a rather dark scene with a high contrast created due to the local light sources

Also the cameras properties are subject to domain randomization. In contrast to other approaches, we only randomize the camera z-position to vary the distance from camera to object. Furthermore, we randomize the focus distance, f-stop and aperture blades in order to generate images with random out-of-focus blur.

3.4 Annotations

One of the main advantages of synthetic image generation is that creating annotations can be performed automatically. In addition to rendering images, we also generate a variety of annotations useful for different computer vision tasks. Our system currently generates ground truth data in the form of bounding boxes, segmentation masks as well as depth maps (as illustrated in Fig. 5).

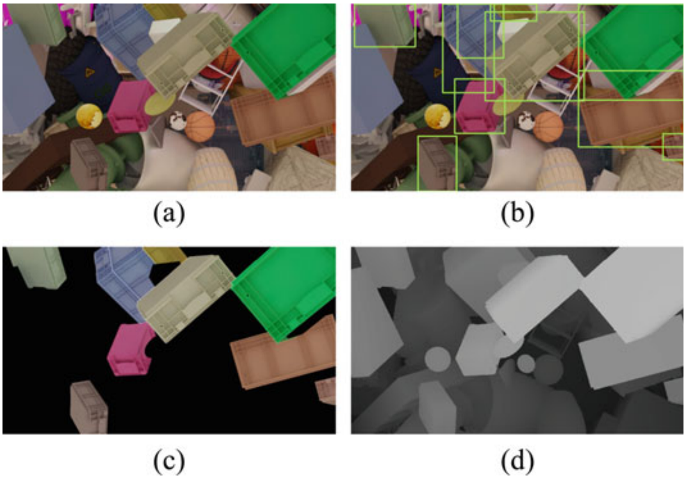
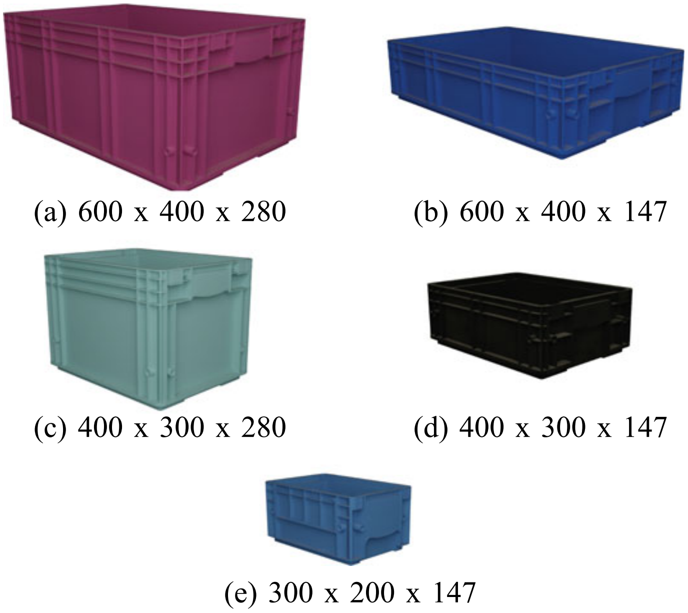


Fig. 5 Ground-truth generation. For each rendered image (**a**), our system automatically generates bounding boxes (**b**), panoptic segmentation masks (**c**) as well as depth maps (**d**)

4 Experiments

We now present experiments of an object detection model trained only on synthetically generated images using the proposed approach. Furthermore, an ablation study was conducted and provides detailed insights into the design principles for reducing the domain gap featured in this novel approach.

In order to do so, we chose a single object class, namely small load carriers (SLCs), because they are widely used in industry, they are standardised and are also subject to automated material flow handling by robots in the future. We believe that synthetic image generation techniques for use in logistics could allow such applications. In sum, SLCs are used for transporting materials and are standardized by the German Association of the Automotive Industry (VDA). Furthermore SLCs are available in different sizes as illustrated in Fig. 6 which can be consolidated and stacked fitting onto a pallet (1.200 mm × 800 mm). In the context of our experiments, we consider load carrier types VDA RL-KLT 3147, 4147, 4280, 6147 and 6280.


(a) 600 x 400 x 280

(b) 600 x 400 x 147

(c) 400 x 300 x 280

(d) 400 x 300 x 147

(e) 300 x 200 x 147

Fig. 6 Small load carriers and their respective size (length × width × height) in millimeters. VDA RL-KLT 6280 (**a**), 6147 (**b**), 4280 (**c**), 4147 (**d**) and 3147 (**e**) are standardized by the Association of the Automotive Industry (VDA) and can be found in logistics throughout different sectors

4.1 Industry-Centered Evaluation Dataset

In order to investigate the different characteristics of our method in a systematic and controlled way an evaluation dataset with natural images was created (see Fig. 7). The evaluation dataset was recorded at the chair's research facility, resembling a realistic industrial environment. Hereby special emphasis was put on various aspects of capturing known influences in the industrial environment. The evaluation dataset contains images under different lighting conditions, with varying number of objects, different object sizes, different distances to the target objects as well as logistically specific states of the SLCs such as loaded as well as stacked load carriers. In total, the dataset consists of 1460 manually annotated images, which can be divided into seven categories:

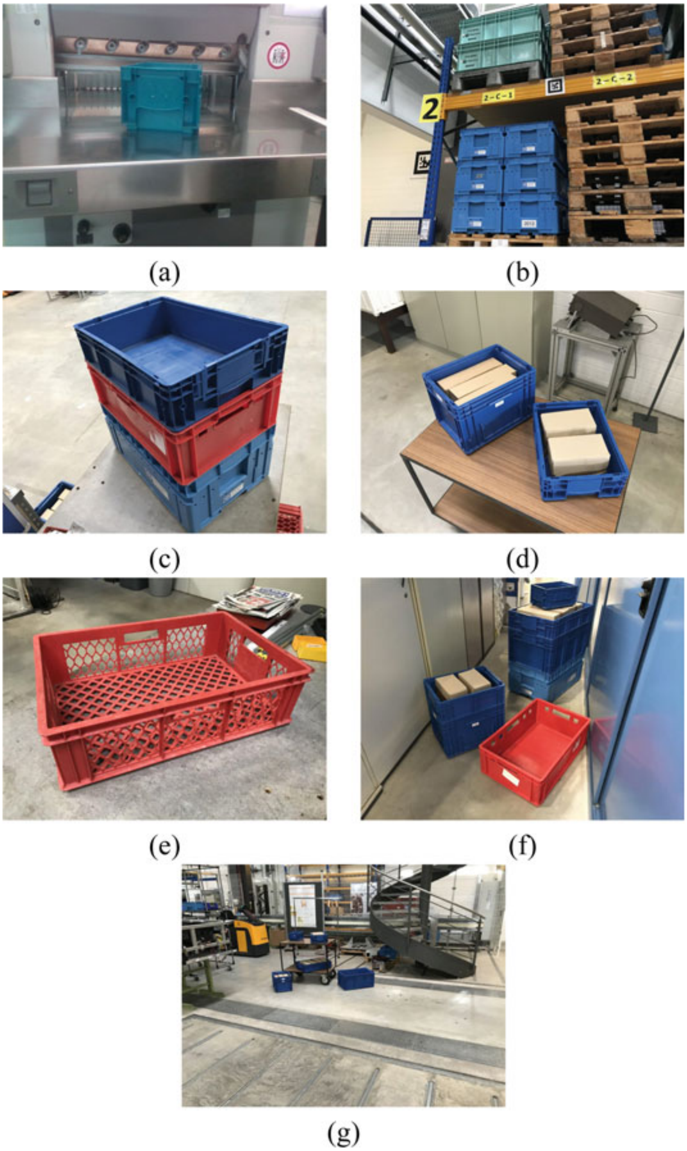
- single-object images
- multi-object images
- images with small object instances (<1 % image area)
- images with medium object instances (between 1 and 10 % image area)
- images with large object instances (>10 % image area)
- loaded SLCs
- stacked SLCs.

4.2 Model Training

We evaluate our method by training a deep neural network object detector. The darknet framework and in particular the YOLOv3 model [35] was chosen for object detection. The model is trained exclusively on synthetic data and evaluated on natural test images. For training a pretrained feature extractor is used. From a network structure perspective, only the number of convolutional filters is adjusted to accommodate the single object category. Other parameters as well as the augmentation strategy remain unchanged. The model was trained on a Nvidia V100 GPU.

4.3 Evaluation and Ablation Study

Finally we describe the evaluation process and ablate different design choices within our image generation approach. Unless otherwise noted, the training was performed as described in Sect. 4.2. We present the precision as well as recall metric for each of the models at an intersection over union (IoU) of 0.5.



(a) (b) (c) (d) (e) (f) (g)

Fig. 7 Industry-centered evaluation dataset. In order to evaluate our approach, we captured and annotated a dataset containing images with single (**a**), multiple (**b**), stacked (**c**) and loaded (**d**) small load carriers. Furthermore, the images can be distinguished by the relative size of the SLC in an image (large (**e**), medium (**f**) and small (**g**)). Note that the classification is not mutually exclusive

4.3.1 Render Engines

Firstly, we compared the effects of images generated using Cycles and EEVEE, two different render engines within Blender. Cycles renders images by tracing back light paths and accumulating them, causing it to be slow, but rather physically correct. EEVEE in contrast is a game-engine and creates images by projecting images from the 3D space to a 2d plane. As these projections do not take into consideration lighting and tracing lightrays, these images look less realistic, but are also generated faster. Figure 8 presents the test results of YOLOv3 trained on 2.500 images generated by Cycles and EEVEE respectively. Trained on the same amount of images, it can be concluded that the detector trained on images rendered by Cycles outperforms the one that was trained on EEVEE-rendered images. Due to the fact that EEVEE ignores the physical realism, it can render images much faster than Cycles. In our experiments, the time for Cycles to generate 2.500 images is equal to the time for EEVEE to generate 5.700 images using devices with same computational capabilities. Therefore, we further investigated the effect of 2.500 Cycles images and 5.700 EEVEE images. It can be seen that increasing the size of the training set improves detection performance. In conclusion, it can be summarized, that the information content per synthetic image generated using Cycles is higher. In contrast, EEVEE is able to generate more images in the same amount of time, but these images contain less information useful for training the network. We chose to continue our experiments with Cycles as a render engine (denoted as standard in following figures).

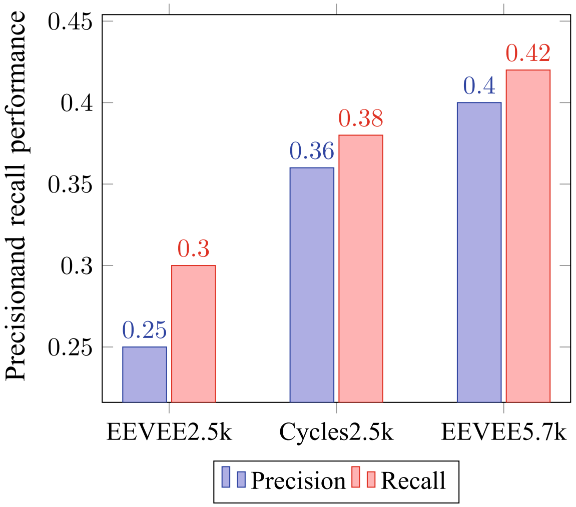


Fig. 8 Ablation study on render engines. Detection performance of a model fully trained on synthetic images rendered by Cycles and EEVEE, two different render engines within Blender

4.3.2 Background Objects and Distractors

In the next experiment we evaluated the importance of background objects and distractors in our approach. To do so, two additional image batches were created, one without background objects (but the background plane), and another one without distractors. Figure 9 illustrates our findings and compares them to the performance of the standard model with background objects and distractor. It can be observed that precision and recall decrease slightly after removing the background objects. This suggests that the background objects are not as effective as in experiments presented by [21]. Besides, this experiment clearly shows the importance of using distractors as the detection performance plummets when removing them.

4.3.3 Object Relation Modelling

In order to study the effects of ORM in our approach, we used the standard image batch with a single modelled relation and generated two additional image batches; one batch without modelled relations (i.e. objects were placed randomly), and one batch with multiple different relations. Again, the trained detectors were tested on natural images, many of which contain SLC stacks and loaded SLCs. Figure 10 suggests, that increasing the number of applied relations in synthetic training images improves the performance of the detector. Furthermore, we tested these detectors on two different subsets of our dataset. One subset containing only images with SLC stacks ("stacks"

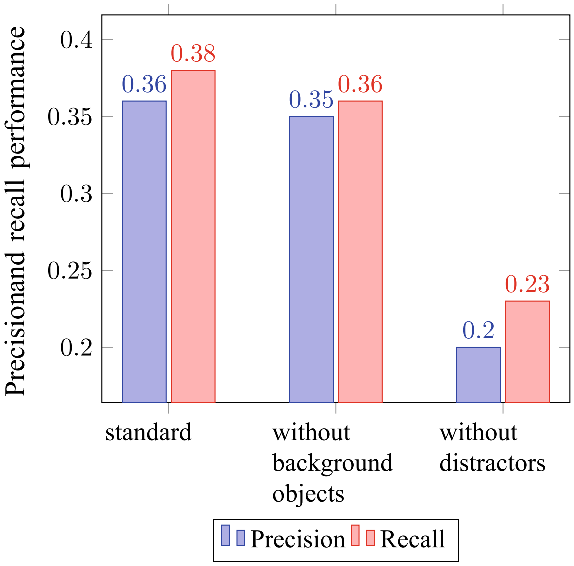


Fig. 9 Ablation study on background objects and distractors. Detection performance of a model fully trained on synthetic images, synthetic images without background objects and synthetic images without distractors

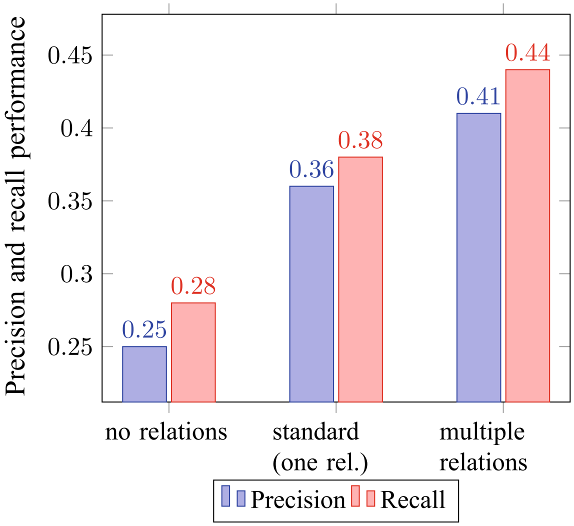


Fig. 10 Ablation study on Object Relation Modelling (ORM). Detection performance of a model fully trained on synthetic images with one modelled relation (standard), synthetic images without ORM and synthetic images with multiple modelled relations

subset) and another subset containing images of loaded SLCs ("loaded" subset). The results are shown in Fig. 11 and indicate a similar tendency as shown before. Furthermore, this suggests that ORM has similar effects on both subsets.

4.3.4 Object Size

Finally, we performed experiments to investigate how the detector trained on synthetic images generated by our method performs on real test images with respect to objects of different size. The results presented in Fig. 12 show that the performance drops prominently as the object size decreases. When detecting large objects, the precision reaches up to 0.81. However, for the challenging detection of small objects, it drops to only 0.11.

5 Conclusion

In this paper we have presented a scalable approach for synthetic image generation of industrial objects utilizing well-working features and augmenting them with novel components, such as Object-alike Distractors (OAD), Object Relation Modelling

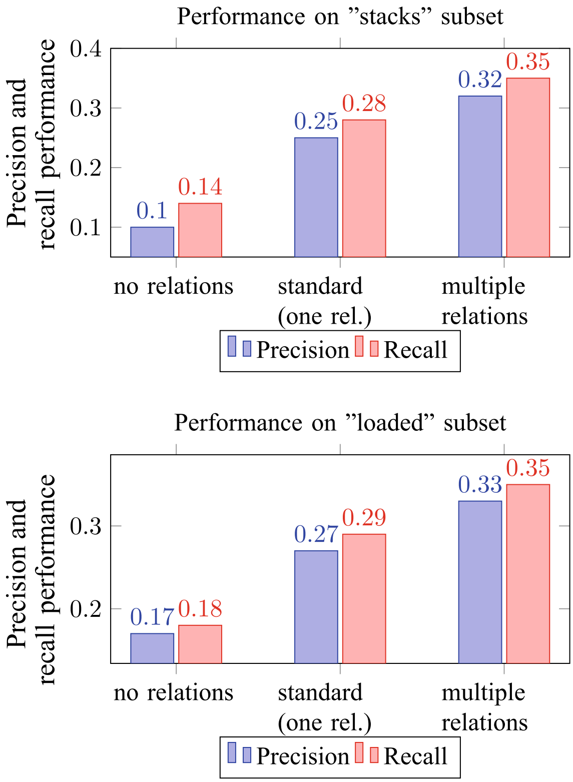


Fig. 11 Ablation study on Object Relation Modelling (ORM) focussing on "stacks" (top) and "loaded" (bottom) subset containing images where understanding the relation between objects is necessary. For each subset, we present the detection performance of a model fully trained on synthetic images with one modelled relation (standard), synthetic images without ORM and synthetic images with multiple modelled relations

(ORM) and a Mixed-lighting Illumination (MLI). Due to missing industrial datasets, we generated and presented a industry-centered dataset for evaluation purposes of synthetic image generation methods. Finally, an extensive ablation study is presented, wrapping up our experiments.

Most importantly, we show, that our approach enables fully synthetic training for object detection in industry. Still, in its current state, it is limited to certain boundaries and the detection performance is worse compared to detectors trained on natural images only. Furthermore, we show that our novel features (Object Relation Modelling, Object-alike Distractors and Mixed-lighting Illumination) are simple, effective and scalable (i.e. are able to be automated) methods to consider when developing synthetic image generation methods. They work by changing the statistics of certain features within the synthetic dataset and 'guiding' the detection model to focus on these features. All of this, whilst still being scalable, without the need of

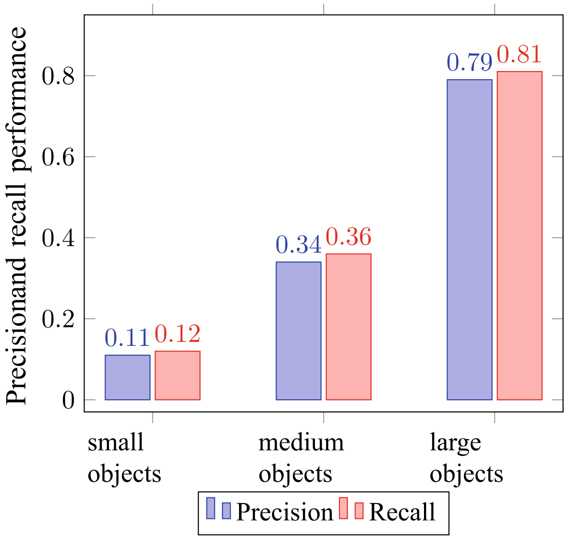


Fig. 12 Ablation study on object size. Detection performance of a model fully trained on synthetic images with one relation (standard) and synthetic images with more relations tested on subsets of our dataset containing large, medium and small objects

manually modelling different 3D scenes. Finally, we found that backwards ray-traced rendering increases the information entropy within synthetic datasets, suggesting that physical correctness is important for current convolutional neural networks.

For future research we plan to extend our approach and further analyse Blenders capabilities as a synthetic image generation system. This requires additional experiments with different object classes in diverse industrial applications to prove the generalization and robustness of our method. Furthermore, we will be expanding our tests to other network architectures as well as computer vision tasks in order to investigate the domain gap in different tasks. Finally, the domain gap is still apparant, so future research needs to focus on minimizing it.

Acknowledgements We thank the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW) for the provision and support of cloud computing infrastructure essential to this publication. All support is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the writers and do not necessarily reflect the views of the BMW Group or the LRZ.

References

1. S. Ren, K. He, R.B. Girshik, J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* (2015)
2. J. Redmon, A. Farhadi, *Yolo9000: Better, Faster, Stronger* (2016)

3. T. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollár, *Focal Loss for Dense Object Detection* (2017)
4. M. Tan, R. Pang, Q.V. Le, *EfficientDet: Scalable and Efficient Object Detection* (2019)
5. K. He, G. Gkioxari, P. Dollár, R.B. Girshick, *Mask R-CNN* (2017)
6. W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, Z. Wang, *FasterSeg: Searching for Faster Real-Time Semantic Segmentation* (2020)
7. Z. Su, M. Ye, G. Zhang, L. Dai, J. Sheng, *Cascade Feature Aggregation for Human Pose Estimation* (2019)
8. A. Bulat, J. Kossaifi, G. Tzimiropoulos, M. Pantic, *Toward Fast and Accurate Human Pose Estimation Via Soft-Gated Skip Connections* (2020)
9. C. Godard, O. Mac Aodha, G.J. Brostow, *Unsupervised Monocular Depth Estimation with Left-Right Consistency* (2016)
10. J.H. Lee, M. Han, D.W. Ko, I.H. Suh, *From Big to Small: Multi-scale Local Planar Guidance for Monocular Depth Estimation* (2019)
11. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, *Vision Meets Robotics: The KITTI Dataset* (2013)
12. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, D. Anguelov, *Scalability in Perception for Autonomous Driving: Waymo Open Dataset* (2019)
13. M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, *The PASCAL Visual Object Classes (VOC) challenge* (2010)
14. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, *Microsoft coco: Common Objects in Context* (2014)
15. H. Noh, A. Araujo, J. Sim, B. Han, *Image Retrieval with Deep Local Features and Attention-Based Keypoints* (2016)
16. F.E. Nowruzi, P. Kapoor, D. Kolhatkar, F. Al Hassanat, R. Laganiere, J. Rebut, *How Much Real Data do we Actually Need: Analyzing Object Detection Performance Using Synthetic and Real Data* (2019)
17. X. Peng, K. Saenko, *Synthetic to Real Adaptation with Deep Generative Correlation Alignment Networks* (2017)
18. D. Dwibedi, I. Misra, M. Hebert, *Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection* (2017)
19. M. Johnson-Roberson, C. Barto, R. Mehta, S.N. Sridhar, K. Rosaen, R. Vasudevan, *Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks?* (2017)
20. A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, S. Birchfield, *Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data* (2018)
21. S. Hinterstoisser, O. Pauly, H. Heibel, M. Marek, M. Bokeloh, *An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance an Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection* (2019)
22. X.B. Peng, M. Andrychowicz, W. Zaremba, P. Abbeel, *Sim-to-Real Transfer of Robotic Control with Dynamics Randomization* (2018)
23. M. Rad, M. Oberweger, V. Lepetit, *Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images* (2018)
24. J. Borrego, A. Dehban, R. Figueiredo, P. Moreno, A. Bernardino, J. Santos-Victor, *Applying Domain Randomization to Synthetic Data for Object Category Detection* (2018)
25. X. Pan, P. Luo, J. Shi, X. Tang, *Two at Once: Enhancing Learning and Generalization Capacities Via ibn-net* (2018)
26. G. Yang, H. Xia, M. Ding, Z. Ding, *Bi-directional Generation for Unsupervised Domain Adaptation* (2020)
27. S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, K. Bousmalis, *Sim-to-Real via Sim-to-Sim Data-Efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks* (2019)

28. S. Thalhammer, K. Park, T. Patten, M. Vincze, W. Kropatsch, *Sydd Synthetic Depth Data Randomization for Object Detection Using Domain-Relevant Background* (2019)
29. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World* (2017)
30. N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, T. Brox, What makes good synthetic training data for learning disparity and optical flow estimation? Int. J. Comput. Vis. **126**(9), 942–960 (2018)
31. J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, *Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization* (2018)
32. B. Mehta, M. Diaz, G. Florian, C. J. Pal, L. Paull, *Active Domain Randomization* (2019)
33. I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, L. Zhang, *Solving Rubik's Cube with a Robot Hand* (2019)
34. Y. Movshovitz-Attias, T. Kanade, Y. Sheikh, *How Useful is Photo-Realistic Rendering for Visual Learning?* (2016)
35. J. Redmon, A. Farhadi, *Yolov3: An Incremental Improvement* (2018)

Research on the Prediction of Shenzhen Growth Enterprise Market Price Index Based on EMD-ARIMA Model

Tianhua Li, Shaowei Qu, and Gaoping Huang

Abstract Under the background of mass entrepreneurship and innovation, it is of great practical and theoretical significance to predict the Shenzhen Growth Enterprise Market (GEM) price index. In view of the nonlinearity and the non-stationarity of GEM price index a prediction method based on empirical mode decomposition (EMD) and Autoregressive Integrated Moving Average (ARIMA) model is studied. Firstly, the EMD is used to stabilize GEM price index data to make GEM price index data more regular and improve the non-linear and non-stationary characteristics of GEM price index data. Then, the ARIMA model is used to model and predict the decomposed data. The model accuracy evaluation index results of the EMD-ARIMA model in this paper are lower than that of the ARIMA model and the average error rate of the prediction results is also lower than the ARIMA model. The results show that the method proposed in this paper is more accurate than that of direct prediction by only using ARIMA model, which indicates that the EMD-ARIMA method proposed in this paper has higher generalization ability and prediction accuracy.

Keywords Data mining · Time series · The growth enterprise market price index · EMD-ARIMA

T. Li (✉) · G. Huang
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: rongren5064@sina.com

G. Huang
e-mail: mengtianhgp@163.com

S. Qu
School of Humanities and Social Sciences, University of Science and Technology Beijing, Beijing, China
e-mail: 13810606721@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_54

1 Introduction

The China's Shenzhen Growth Enterprise Market (GEM) officially opened in Shenzhen Stock Exchange on October 30, 2009. The GEM is different from the main board market, most of the companies listed on GEM are mainly high-tech businesses, which have a high growth rate. At the same time, most of the stock price earnings ratio of GEM is higher than that of the main board market. The GEM market is a relatively high risk and low listing threshold, which often leads to the sharp rise and fall of GEM price index, with the characteristics of high risk and high return. Under the background of mass entrepreneurship and innovation, talents with high and new technology will be attracted to start their own businesses and set up high-tech companies. At the same time, as the GEM registration system has been implemented, more high-tech companies will choose to list on the GEM. The effective prediction of the GEM price index is the key link for investors to increase returns and avoid risks in investing in GEM stocks. Only through effective time series prediction and mining of the GEM price index data can we better reduce the losses caused by the sharp fluctuations of the GEM price index and at the same time promote the healthy and sustainable development of the GEM. The effective forecasting and modeling of the GEM price index is of great significance and an important subject studied by many scholars. Therefore, it is necessary to use the appropriate model method to predict the GEM price index.

In recent years, due to the good nonlinear fitting ability of data mining technology, it is often used for stock price prediction. The commonly used methods include support vector machine method [1–3], random forest method and artificial neural network method [4, 5]. The Support Vector Machines (SVM) have great advantages in processing small samples, but the selection of some parameters will directly determine the accuracy of the prediction. The random forest method has a strong generalization ability, but the random forest is a typical black box model, so the prediction results are not easy to explain. Similarly, the artificial neural network is a black box model with complex structure and poor interpretability. However, the GEM price index data has significant non-stationary, non-linear and highly complex features, which has become a difficult problem for repeatable feature extraction in its prediction. Therefore, even using random forest, ANN, SVM and other new nonlinear machine learning methods is still difficult to fully and effectively explain and predict the prediction results. The stock price index is often used to establish a time series analysis model based on its historical data to predict future data. The Autoregressive Integrated Moving Average (ARIMA) model is often used to predict various time series [6–9]. In this paper, the ARIMA model is selected to predict the GEM index. Considering that the GEM price index data is non-linear and non-stationary, it needs to be stabilized first. Since EMD has more advantages in extracting trend items and can be applied to financial time series prediction [10, 11], this paper selects EMD to stabilize the GEM index price data. This paper draws on the idea of "decomposing before integration" to solve complex system problems. Firstly, the complex system is decomposed by EMD, and then modeled and predicted by mode reconstruction on the basis decomposition. Finally, the predicted values of IMFs and trend term are integrated to obtain the original series predicted values with high accuracy.

For the GEM price index of nonlinear and non-stationary characteristics of financial time series price data, this paper adopts the combination of EMD and ARIMA model to predict the GEM index. In this paper, the GEM price index data comes from NetEase finance and economics, and the time series is modeled and predicted from the daily closing price data of solstice on June 1, 2010 and February 25, 2020. Firstly, the EMD method is used to conduct modal decomposition of the GEM price index data to obtain several decomposed subsequences. Because these subsequences can reflect the components of the original signal at different scales, the decomposed data is more stable than the original data. Second, the decomposed sequences were reconstructed into three sequences of high frequency, low frequency and residual. Then, the ARIMA model is used to model and predict each reconstructed subsequence separately. Finally, the prediction results of all the reconstructed high frequency, low frequency and residual sequences are added to obtain the prediction of the original the GEM price index data.

2 Related Work

A. *The EMD-ARIMA Model Framework*

The steps for building the composite model are as follows:

Step1: the EMD method is used to conduct modal decomposition of the GEM price index data to obtain *n* IMFs and one residual trend item.
Step2: reconstruct IMFs to obtain high frequency, low frequency and residual terms, and establish ARIMA prediction model respectively.
Step3: calculate the predicted values of high frequency, low frequency and residual trend terms respectively by using ARIMA prediction model, and then add and integrate to get the final forecast value of the GEM price index.
Step4: evaluate the EMD-ARIMA and ARIMA model with various evaluation indexes. then, compare the predicted value error of EMD-ARIMA model with the predicted value error of ARIMA model. finally, draw a conclusion.

The specific model framework of this paper is shown in Fig. 1.

B. *The EMD model decomposition algorithm*

The Empirical Mode Decomposition (EMD) is a Chinese American Norden E Huang by working in NASA scientists and its academic partners and jointly present in 1998 [12], is a fundamentally different from the traditional methods such as Fourier transform new time-frequency analysis method, the role of is considered to be NASA's most important inventions in the field of mathematics application of the theory. As an important model, the doctoral student Yu of the Chinese academy of sciences used it to predict oil futures prices and achieved good results.

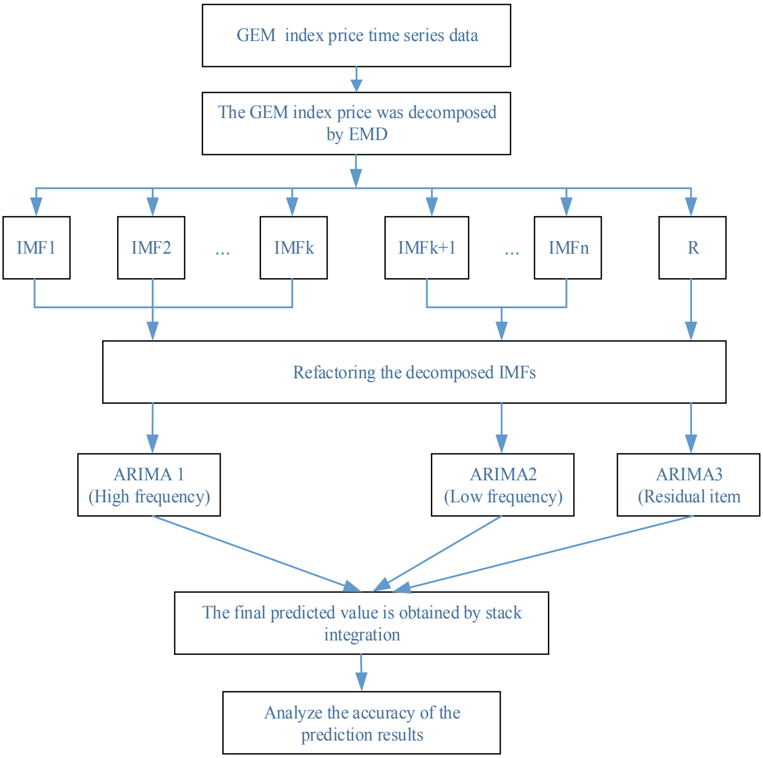

GEM index price time series data
The GEM index price was decomposed by EMD
IMF1
IMF2
...
IMFk
IMFk+1
...
IMFn
R
Refactoring the decomposed IMFs
ARIMA 1 (High frequency)
ARIMA2 (Low frequency)
ARIMA3 (Residual item
The final predicted value is obtained by stack integration
Analyze the accuracy of the prediction results

Fig. 1 Schematic diagram of the framework of EMD-ARIMA model

The basic principle of Empirical Mode Decomposition (EMD) method is to decompose different characteristic scales or trends in the signal step by step to produce a limited number of inherent modal functions (IMF). The decomposed IMF component reflects the components of the original signal with different scale fluctuations. Therefore, EMD is an effective method to deal with nonlinear signals. The specific process of its decomposition is as follows:

Step1: find out all maximum points of $X(t)$ and use interpolation function to fit them into the upper envelope of the original sequence. Similarly, the lower envelope of the original sequence is fitted.
Step2: calculate the mean value of upper and lower envelope, denoted as $m1$.
Step3: calculate $h1 = X(t) - m1$, $h1$ is the new sequence.
Step4: judge whether $h1$ meets the two characteristics of IMF, that is: (1) the number of extremum points and zero crossing is the same or the maximum difference is 1; (2) at any time, the average of its upper and lower envelope must be, so $h1$ is an IMF; Continue with step 5. If not, calculate the upper and lower envelope of $h1$ and repeat steps 2, 3 and 4 until IMF conditions are met.

Step5: subtract the extracted IMF from $X(t)$, repeat steps 1–4 to obtain the next IMF until all the IMF are extracted. Thus, $X(t)$ can be written as:

$$X(t) = \sum_{i=1}^{n} C_i(t) + R_n(t), \quad i = 1, 2, \ldots, n \tag{1}$$

where, n is the number of IMF, $R_n(t)$ is the final residual term, which generally represents the long-term trend of the sequence *X(t)*, and $C_i(t)$ is the IMF component of each layer, including the local variation characteristics of the sequence.

C. *Autoregressive Integrated Moving Average (ARIMA) model*

ARIMA model is composed of AR, I and MA, where AR and MA are autoregressive model and moving average model respectively, while I represents single-order integer, that is, the difference order converted to stationary sequence. Generally speaking, the differential integration moving average autoregressive model is developed on the basis of the autoregressive moving average model (ARMA). The ARIMA model is formed by the ARMA model for the data after d-order difference, and it can effectively predict the univariate stationary time series. The ARIMA(p, d, q) model is:

$$\begin{cases} w_t = \nabla^d x_t, \\ w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + \varepsilon_t w_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \\ \phi_p \neq 0, \phi_q \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, \\ E(\varepsilon_t \varepsilon_s) = 0, s \neq t; E(x, \varepsilon_t) = 0, \forall s < t. \end{cases} \tag{2}$$

where, d is the difference order, p is the autoregressive order, and q is the moving order.

The specific ARIMA model flow diagram proposed in this paper is as shown in Fig. 2.

Step1: input the time series data and verify the stationarity of the sequence. By observing the sequence diagram and ADF unit root test, we can judge whether the sequence is stable.

Step2: For the non-stationary sequence, the difference operation can be used to make it stable, in which the difference degree d (or the automatic algorithm can be directly used to determine the parameters of p, d and q in the model).

Step3: and then carry out white noise test on the stationary sequence after the difference d times. If the stationary sequence is non-white noise sequence, step 4 can be entered.

Step4: The model identification and order determination. According to the autocorrelation graph and partial autocorrelation graph of the sample, the appropriate

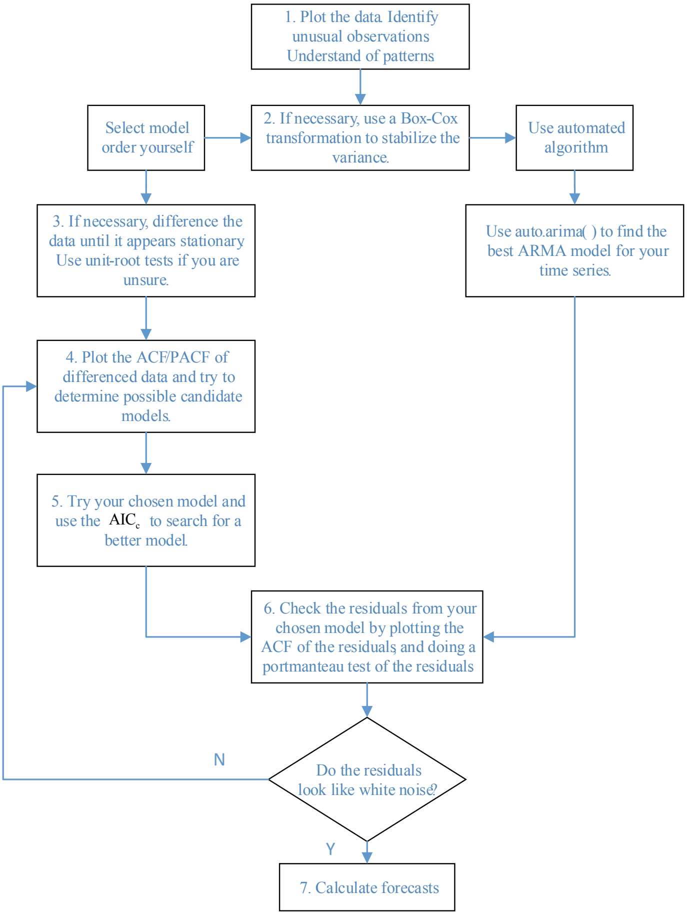

1. Plot the data. Identify unusual observations Understand of patterns
Select model order yourself
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Use automated algorithm
3. If necessary, difference the data until it appears stationary Use unit-root tests if you are unsure.
Use auto.arima( ) to find the best ARMA model for your time series.
4. Plot the ACF/PACF of differenced data and try to determine possible candidate models.
5. Try your chosen model and use the AIC$_c$ to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals
N
Do the residuals look like white noise?
Y
7. Calculate forecasts

Fig. 2 General process for forecasting using an ARIMA model

autocorrelation order p and the moving average order q were selected, and then the model was fitted. If the autocorrelation coefficient decays to zero exponentially, then there is an autoregressive (AR) process. If the partial autocorrelation coefficient decays to zero exponentially, there is a moving average (MA) process. If both of the above conditions exist, this indicates that the sequence is a summation Autoregressive Integrated Moving Average model (ARIMA) process.

Step5: parameter estimation of the model. The least square method is used to estimate the regression coefficient of the sequence. The least square method can make full use of the information of the observed value of the sequence and improve the prediction accuracy of the sequence.

Step6: test the model. If the residual sequence is not white noise sequence, return to step 3 and rebuild the model until it passes the parameter test and model residual white noise test.

Step7: prediction of the model. The ARIMA model, which passed the parameter test and residual white noise test, was used for prediction.

D. *Prediction Result Evaluation Method*

In this paper, the ME, RMSE, MAE and MASE were used to measure the prediction effect of ARIMA and EMD-ARIMA fusion models. $\hat{y}_i$ and y_i are used to represent the predicted value and the true value respectively. The smaller the index value is, the better the prediction effect is.

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{y}_i - y_i\right)\right| \tag{5}$$

$$\mathrm{MASE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{\left|\left(\frac{1}{(n-1)}\sum_{i=2}^{n}|y_i - y_{i-1}|\right)\right|} \tag{6}$$

3 Result Analysis

A. *Data Preparation*

This paper selects solstice on June 1, 2010 and GEM daily price index data on February 18,2020 from NetEase finance to establish the time series prediction model. The original GEM price index data is shown in Fig. 3.

B. *EMD—ARIMA composite model of GEM price index forecast modeling process*

Based on EMD-ARIMA model portfolio of price index to predict the basic process of the GEM as follows. Firstly, using the empirical mode decomposition of the original time series is decomposed into more stable sequence. Then, using ARIMA model modeling projections for each subsequence, predicted by the modal component may be all forecasts, put together. Finally, the prediction results of GEM price index is obtained.

The specific steps are as follows:

Step1: EMD method is used to decompose the first 2640 original data into IMF components and residual terms of several different feature scales.

Step2: reconstruct IMF to obtain high frequency, low frequency and residual sequences. ARIMA model was used to model 2640 high frequency, low frequency and residual sequence points.

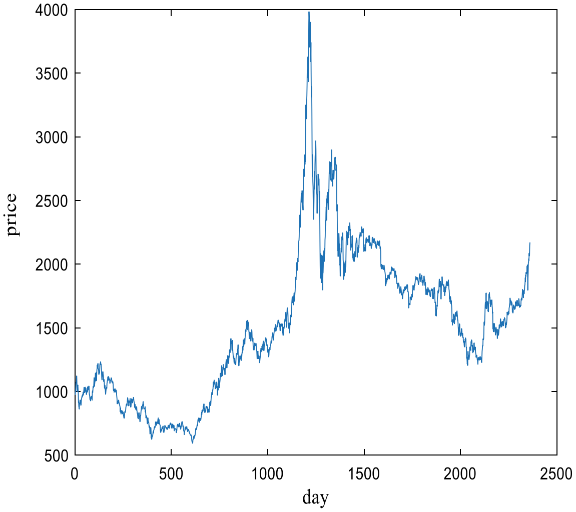


Fig. 3 The original GEM price index data trend time series line chart

Step3: the ARIMA model is used to predict each high frequency, low frequency and residual terms, and then all the predicted results are added up to obtain the predicted value of GEM index sequence data.

C. *EMD decomposition and integration of GEM price index sequence*

EMD decomposition of GEM price index sequence. In this paper, 7 IMFs and one residual term are obtained by decomposing the EMD of the obtained GEM price index data. As shown in the following figure, IMF1-IMF7 is arranged from high frequency to low frequency and has a residual term. As can be seen from the figure, all IMF frequencies change with time, and the frequencies decrease in sequence. IMF1 represents the high-frequency feature of the sequence, IMF7 represents the low-frequency feature of the sequence, and the residual term represents the long-term average trend of the sequence. In fact, EMD decomposition is the decomposition of a nonlinear, non-stationary sequence into several components of different frequencies and a long-term average trend term, as shown in Fig. 4.

The EMD restructuring. The mean of each IMF was calculated, the first mutation point significantly deviated from 0 was found, and the secondary mutation point and all previous IMF reorganizations were added into a new high-frequency sequence. Restructuring the rest of the IMF into new low-frequency sequences. The residual term is trend term. In this paper, IMF1-IMF4 is formed into a high-frequency sequence, and IMF5-IMF7 into a low-frequency sequence. The residual term constitutes the trend term. The mean of each IMF components shown in Fig. 5 and Fig. 6.

D. *EMD-ARIMA model accuracy evaluation and prediction performance analysis*

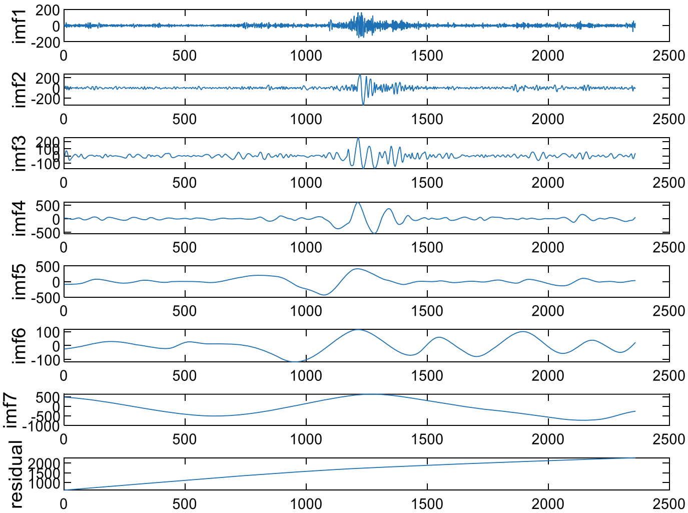


Fig. 4 EMD decomposition results

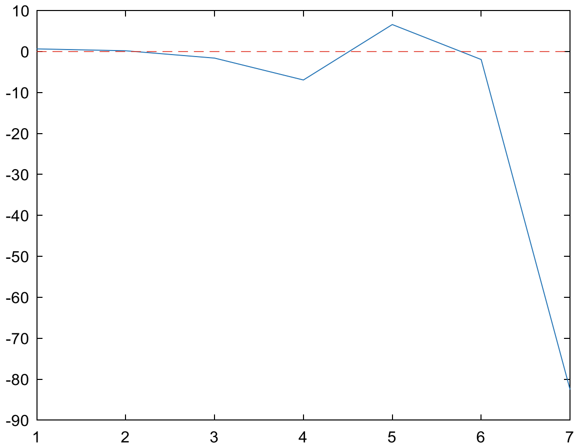


Fig. 5 The mean of each IMF component

In order to evaluate the accuracy of the time series prediction model constructed in this paper, it is necessary to select appropriate evaluation indicators. This paper uses the evaluation indicators of ME, RMSE, MAE, and MASE regression methods to quantitatively evaluate the fitting effects of the EMD-ARIMA and ARIMA models. The corresponding index measurement results are shown in Table 1.

In this paper, the mean error (ME), mean absolute error (MAE), root mean square error (RMSE) and mean absolute standardization error (MASE) are used to measure

Fig. 6 GEM price index data EMD after the reconstruction

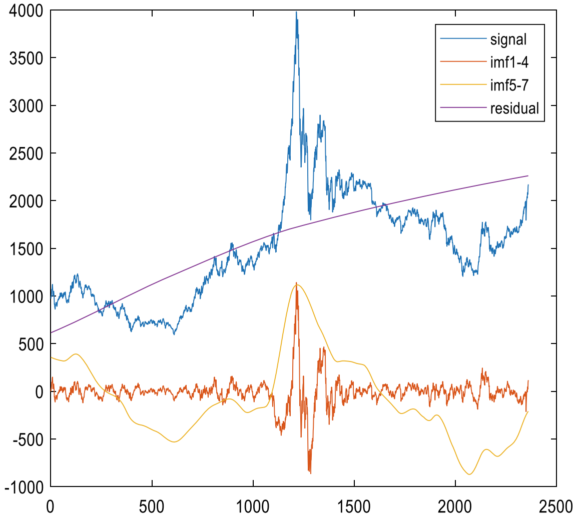


Table 1 ARIMA model and EMD-ARIMA error evaluation index results

Model type	Evaluation index			
	ME	RMSE	MAE	MASE
ARIMA	0.325	36.297	22.656	0.999
EMD-ARIMA	0.006	11.830	7.402	0.330

the prediction error. It can be seen that the error index of EMD-ARIMA combination method is significantly lower than that of ARIMA model alone, which also indicates that EMD-ARIMA model has stronger generalization ability.

In order to verify the effectiveness of the method in this paper, the constructed ARIMA model and EMD-ARIMA model are respectively used to predict the GEM price index data in the next 20 days, and the predicted results are compared with the actual GEM price index data and the average forecast error rate is calculated. The prediction results of the two prediction models are shown in Fig. 7 and Fig. 8. As can be seen from Fig. 7 and Fig. 8, the overall fitting result of the combined model is better than that of the ARIMA model alone.

From two kinds of prediction error rate in Fig. 9, the EMD-ARIMA model to predict the effect is better than using ARIMA model prediction effect. ARIMA model to predict the results of the average error rate is 3.54%, the EMD-ARIMA predictive results of the ARIMA model portfolio average error rate was 2.65%. The average error rate of the EMD-ARIMA model is lower than that of the ARIMA model, indicating that the prediction accuracy of the combined model has been significantly improved. At the same time can also see that the longer the combination model for their future prediction error rate significantly lower than that of ARIMA model prediction error rate, combination model have more advantages in the long-term forecast, because

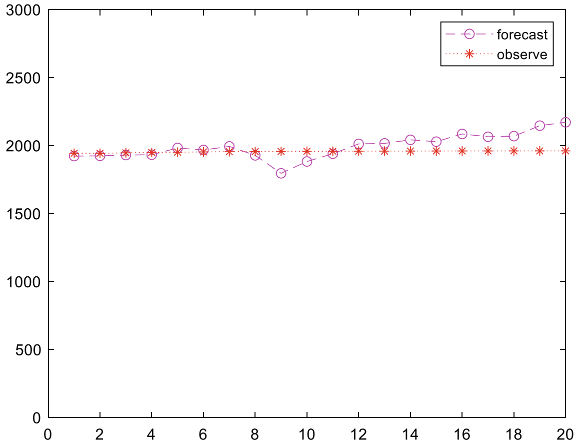


Fig. 7 Based on the ARIMA GEM price index forecast value and the actual value of the comparison

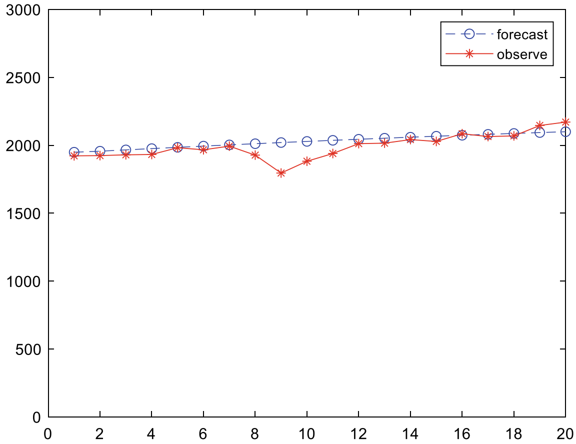


Fig. 8 Based on the EMD-ARIMA GEM price index forecast value and the actual value of the comparison

the EMD can decompose the original stock data into more stable subsequence, better able to find out the inherent law of GEM price index data, so that you can better track using ARIMA model to predict that the actual price index to predict the GEM data.

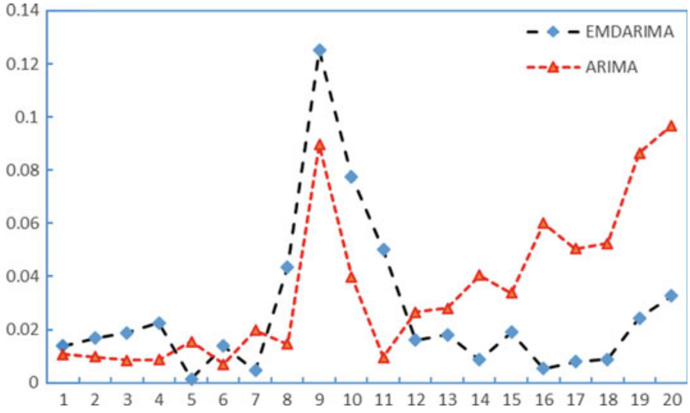


Fig. 9 ARIMA model and EMD-ARIMA composite model for GEM price index forecast error rate results

4 Conclusion

Aiming at the nonlinearity and non-stationarity of gem index price data, this paper studies the combined model prediction method based on empirical mode decomposition (EMD) and autoregressive comprehensive moving average (ARIMA). The following conclusions are drawn. First of all, the EMD method can decompose and recombine different characteristic scales in the GEM price index sequence to make the gem sequence data more stable and regular, and the EMD-ARIMA model accuracy is higher than that of only using ARIMA model. Secondly, EMD smoothed the stock data, which greatly facilitated the process of modeling and predicting stock data with the ARIMA model, making the average accuracy of the prediction results of the EMD-ARIMA model higher than the ARIMA model. At the same time, compared with the ARIMA model, the combined model mines the hidden different scale information of the time series so that it has a better long-term prediction effect. Finally, the combined model combines the advantages of a single model, with strong versatility and accuracy. However, as the stock market is often impacted by external major events, the price index data may fluctuate violently and the error rate of the portfolio model is relatively large, which requires further research in the future.

References

1. R. Rui, D. Wu, T. Liu, Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst. J. **13**(1), 760–770 (2018)
2. Y. Chen, Y. Hao, A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Syst. Appl. **80**, 340–355 (2017)
3. T. Manoj, D. Kumar, A hybrid financial trading support system using multi-category classifiers and random forest. Appl. Soft Comput. **67**, 337–349 (2018)

4. G. Kim, S. Kim, Variable selection for artificial neural networks with applications for stock price prediction. Appl. Artif. Intell. **33**(1),. 54–67 (2019)
5. Z. Lei, L. Wang, Price trend prediction of stock market using outlier data mining algorithm, in *Proceedings of the 2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, Dalian, pp. 93–98, 2015
6. A. Mustafa, A. Ihsan, H. Zada, Forecasting stock prices through univariate ARIMA modeling. NUML Int. J. Bus. ManaGEMent **13**(2), 130–143 (2018)
7. M. Prapanna, L. Shit, S. Goswami, Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. Int. J. Comput. Sci. Eng. Appl. **4**(2), 13–29 (2014)
8. B.U. Devi, D. Sundar, P. Alli, An effective time series analysis for stock trend prediction using ARIMA model for Nifty Midcap-50. Int. J. Data Min. Knowl. Manag. Process **3**(1), 65–78 (2013)
9. A.A. Adebiyi, A.O. Adewumi, C.K. Ayo, Comparison of ARIMA and artificial neural networks models for stock price prediction. J. Appl. Math. (2014)
10. Nava, Noemi, T. D. Matteo and T. Aste, Dynamic correlations at different time-scales with empirical mode decomposition. Phys. A: Stat. Mech. Its Appl. **502**, 534–544 (2018)
11. L. Hong, Decomposition and forecast for financial time series with high-frequency based on empirical mode decomposition. Energy Procedia **5**, 1333–1340 (2011)
12. N.E. Huang, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. A: Math., Phys. Eng. Sci. **454**(1971), 903–995 (1998)

Research on Location and Path Planning of Distribution Center Based on Improved k-Means Clustering Algorithm and Improved Ant Colony Algorithm

Shuihai Dou, Zhou Yao, Xiaotong Shi, and Guanyi Liu

Abstract In the complex logistics terminal distribution link, the needs of user points and the road level between user points have an important impact on the location of the distribution center and vehicle path planning. The traditional k-means clustering algorithm and ant colony algorithm have some problems in the process of site selection and vehicle route planning in the distribution center, such as the initial cluster center selection is random, and the needs of user points are not considered. It cannot meet the actual logistics terminal distribution requirements. In order to solve the problems of traditional k-means clustering algorithm and ant colony algorithm, in the ant colony algorithm, the center of gravity method is introduced to obtain the initial cluster center, and then the new cluster center calculation method is obtained by introducing the user point demand. In the ant colony algorithm, the parameters are improved by introducing the driving speed of the vehicle, and then the new ant probability calculation formula is obtained by introducing the demand of the user point. Finally, a simulation experiment was conducted through MATLAB. Experimental results show that the improved k-means clustering algorithm and ant colony algorithm are feasible and effective in the logistics end distribution link.

Keywords VRP · Distribution center · Clustering algorithm · Ant colony algorithm

S. Dou · X. Shi · G. Liu
School of Mechanical and Electrical, Beijing Institute of Graphic Communication, Beijing, China
e-mail: doushuihai@126.com

X. Shi
e-mail: s1301671697@163.com

G. Liu
e-mail: 15501092112@163.com

Z. Yao (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 13011151625@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_55

1 Introduction

With the rapid development of China's economy and the reform and innovation of science and technology, in particular, people's living standards have been greatly improved, resulting in consumer demand and the promotion of business activities between enterprises, which has caused close The rapid development of e-commerce in the middle of the year has accelerated the rapid development of the express delivery industry. However, offline logistics is needed to ultimately support such a huge number of online orders and capital flow. The vigorous development of e-commerce, the increase in user demand and order business have led to the continued growth in the demand for express logistics, which not only provides opportunities for the development of China's express logistics industry, but also will bring major challenges. At the same time, the construction of basic service facilities in some mountainous areas, border areas, and rural areas and in poor areas is still quite lacking. At present, the logistics level and logistics services in rural areas of China are relatively backward compared to cities. They are characterized by high cost of single-piece logistics, a high proportion of loss of fresh agricultural products, long logistics service time, and the last mile delivery problem cannot be solved. Rural residents Production and living needs are not met. In this context, the location of the distribution center and vehicle routing planning for the express logistics terminal outlets are particularly important.

The distribution center is a logistics node in the logistics supply chain. It implements the distribution process for distributors, retailers and downstream customers. It uses distribution facilities and information system platforms to load, unload, sort, circulate and process the goods handled by the logistics. Support, distribution Route design and distribution transportation methods provide customized distribution services for customers. Distribution center location selection means that in a given area, one or more suitable locations are selected in the area containing all user points to establish a regional distribution center, or one or more points are selected among the user points in the area to expand A distribution center is established, and each distribution center is responsible for the distribution task of the corresponding user point. Therefore, the location selection decision of the logistics distribution center is crucial and has strategic significance, and its research needs to be more and more in-depth. Zhang Yiwei used the center of gravity method and heuristic algorithm to solve the location of multiple distribution centers in the supply chain environment [1]; Chen Dailian of Chongqing Jiaotong University studied the location of the same city express delivery center through the CFLP model [2]; Yang of Shenyang University Dong analyzed the distribution of profits between different logistics nodes by analyzing the distribution between different user points [3]. Gao Wenqian of Shanxi University cited the improved gravity search algorithm to solve the logistics distribution center location problem and provided a new method for the logistics distribution center location problem [4]. Huang Siqi of Chongqing Jiaotong University directed the clustering unit in each user point for the location of multiple logistics distribution centers in the city, using TOPSIS method to sort, and then using clustering algorithm to cluster to get the final distribution center results [5]. Xuexin Bao, Xiangchun Xing Researched the port solid waste green logistics system based on Ahp and evaluated it [6].

On the premise of satisfying many prerequisites, the problem of moving one or more facilities of the vehicle to several geographically dispersed customer points and planning the operation route of the vehicle so that the total cost is the lowest or the total transportation route is the shortest Planning problems for vehicles. This problem belongs to the classic combinatorial optimization problem, which was first proposed by Dantzig and Ramser in [7]. Zhang et al. [8] creatively proposed the modeling and solution scheme of multi-model dynamic vehicle problems, and Ma et al. [9] also based on multi-model research for multi-model and multi-vehicle vehicle route optimization problems through mutated ant colony The algorithm was studied [9]. From the perspective of green logistics, Liu [10] studied the VRP problem in the limited refueling network for the problem of fuel consumption during vehicle transportation. Tang [11] considered the actual situation of road traffic control in the city and studied the vehicle routing problem for urban distribution. Zhang [12] used spark and hadoop big data technology to experiment and verify the application of ant colony algorithm in VRPTW problem. Guo [13] of Beijing Jiaotong University constructed a rural e-commerce express vehicle routing model based on Zhi rural as an example of distribution problems in rural areas.

In summary, the alternative points for the location of the distribution center are particularly important, but most of the time, the alternative points are selected by experts independently, and only the distance from the user point is considered. At the same time, most researches on vehicle path planning focus on the shortest distance [14, 15]. Therefore, in order to solve the problems of traditional clustering algorithm and ant colony algorithm, this paper proposes an improved k-means clustering algorithm and ant colony algorithm. According to the impact of the initial distribution center candidate point, user point demand and actual road, they are introduced separately. The center of gravity method, the required parameters and the improvement of the Eta parameter in the ant colony algorithm are finally simulated by matlab.

2 Traditional k-Means Clustering Algorithm and Ant Colony Algorithm

A. *Traditional k-means clustering algorithm*

K-means algorithm refers to the solution of multiple distribution centers, mainly to classify users according to the similarity between users and each type, and constantly adjust the user category and category search. The example shown in Fig. 1 can prove the concept of a k-means grouping algorithm that solves the problem of multiple distribution center locations.

(1) After determining the number of distribution centers, for all user points $\left\{x^1, x^2, \ldots, x^m\right\}$. Randomly select k points $\{u_1, u_2, \ldots, u_k\}$ of the area

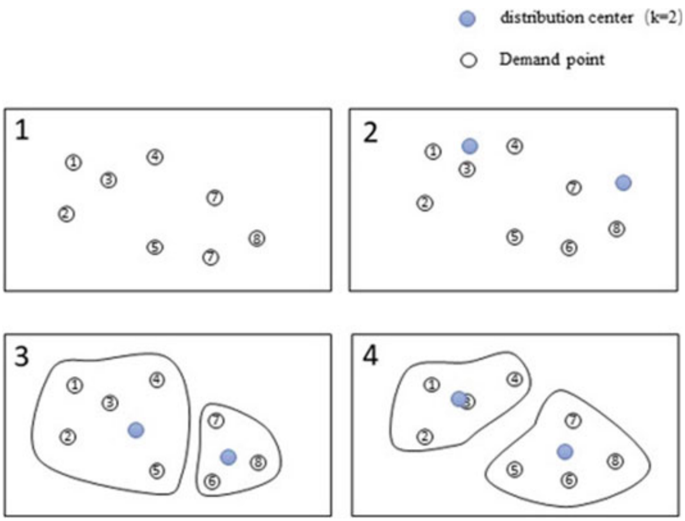


Fig. 1 Schematic diagram of k-mean clustering algorithm

including all user points as the initial cluster center position. Calculate the straight-line distance from each user point to the k distribution centers $[(x^i - x^0)^2 - (y^i - y^0)]^{\frac{1}{2}}$.

(2) According to the distance between each user point and each distribution center, it is divided into the closest distribution center:

$$c^{(i)} = arg\min\left\|x^i - u_j\right\|^2 \tag{1}$$

where $c^{(i)}$ indicates the category corresponding to the i-th user point., x^i represents the coordinates of the i-th user point, u_j indicates the coordinates of the jth distribution center.

According to the newly divided category, calculate the centroid position of the user point in the category, $u_j = (x_i, y_j)$:

$$x_j = \sum_{i=1}^{m} 1\{c^{(j)} = j\}x^{(i)} \Big/ \sum_{i=1}^{m} 1\{c^{(j)} = j\} \tag{2}$$

$$y_j = \sum_{i=1}^{m} 1\{c^{(i)} = j\}y^{(i)} \Big/ \sum_{i=1}^{m} 1\{c^{(i)} = j\} \tag{3}$$

where x_j is the horizontal coordinate of the distribution center in the jth category, y_j is the ordinate of the distribution center in the j-th category.

(3) For the adjusted distribution center, please recalculate the distance from the user point to the distribution center, and then classify it again, and then continuously adjust the location of the distribution center until the classification no longer changes. Finalize the location of the distribution center.

B. *Traditional ant colony algorithm*

The basic ant colony algorithm can be expressed as follows: At the beginning of the algorithm, m ants are randomly placed in n to make a city, and at the same time, the first element of the tabu list of each ant is set to the city where it is currently located. At this time, the amount of pheromone on each path is equal. Assuming $\tau_{ij}(0) = c$ (c is a small constant). Then each ant independently selects the next city based on the amount of pheromone remaining on the path and the heuristic information (the distance between the two cities). At time t, the probability of the ant transferring from city i to city j $p_{ij}^{k}(t)$ can be expressed as:

$$p_{ij}^{k}(t) = \left\{ \begin{array}{cc} \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum\limits_{s \in J_k(i)} [\tau_{is}(t)]^{\alpha}[\eta_{is}]^{\beta}}, & j \in J_k(i) \\ 0 & else \end{array} \right\} \tag{4}$$

where $J_k(i) = \{1, 2, \ldots, n\} - tabu_k$ represents the set of cities that Ant K is allowed to choose next. Tabu Table $tabu_k$ records the city that Ant K currently travels through. When all n cities are added to Tabu Table $tabu_k$, Ant K completes a tour, and the path that Ant K traverses at this time is a feasible solution to the TSP problem [14]. $\eta_{ij}(t)$ is a heuristic factor, indicating the expected degree of ants transferring from city i to city j. In the ant colony algorithm, $\eta_{ij}(t)$ usually takes the reciprocal of the distance between city i and city j. α And β represent the relative importance of pheromones and expected heuristic factors, respectively. After all ants have completed a round trip, the pheromone update on each path can be expressed as:

$$\tau_{ij}(t + n) = (1 - \rho) \cdot \tau_{ij}(t) + \Delta\tau_{ij} \tag{5}$$

where $\rho(0 < \rho < 1)$ represents the evaporation coefficient of the pheromone on the path, $1 - \rho$ represents the persistence coefficient of the pheromone; $\Delta\tau_{ij}$ represents the increment of the pheromone on the edge ij in this iteration:

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{6}$$

where $\Delta\tau_{ij}^{k}$ represents the amount of pheromone that the kth ant stays on the edge i or j in this iteration. If the ant k does not pass through the edge i or j, the value of $\Delta\tau_{ij}^{k}$ is zero. $\Delta\tau_{ij}^{k}$ can be expressed as:

$$\Delta\tau_{ij}^{k} = \left\{ \begin{array}{ll} \frac{Q}{L_k}, & \text{ant k pass through the edge i or j} \\ 0, & \text{else} \end{array} \right\} \quad (7)$$

where Q is a normal number, and L_k represents the length of the path ant k traversed during this round trip.

Ant colony algorithm is actually an algorithm combining positive feedback principle and heuristic algorithm. When choosing a path, ants not only use the pheromone on the path. Moreover, the reciprocal of the distance between cities is used as a heuristic factor.

The specific implementation steps of the basic ant colony algorithm are as follows:

(1) Let time t = 0 and cycle number $N_c = 0$, set the maximum cycle number G, place m ants on n elements, and make the initial information amount of each edge (i, j) on the directed graph $\Delta\tau_{ij} = 0$.
(2) Cycles $N_c = N_c + 1$.
(3) Ant's tabu table index number k = 1.
(4) Number of ants k = k + 1.
(5) The individual ant selects the element j according to the probability calculated by the state transition probability formula and advances.
(6) Modify the tabu list pointer, that is, move the ant to a new element after selection, and move the element to the tabu list of the individual ant.
(7) If the elements in set C are not traversed, that is, k < m, then jump to step (4); otherwise, execute step (8).
(8) Record this best route.
(9) Update the amount of information on each path.
(10) If the end condition is satisfied, that is, if the number of cycles $N_c \geq$ G, the cycle ends and the program optimization result is output; otherwise, the tabu list is cleared and jump to step 2.

3 Improvement of k-Means Clustering Algorithm and Ant Colony Algorithm

A. *Improvement of k-means clustering algorithm*

The k-means clustering algorithm has some shortcomings, but its main disadvantage is that the initial clustering center must be randomly generated before clustering, but most of the time the location of the obtained initial center is unknown, and the final result is Location sensitive.

Therefore, to improve the k-means clustering algorithm, the initial random determination of the initial clustering center is changed to k categories by angles first, and then the user points of the US class are first used to generate the initial clustering center, The specific process is as follows:

(1) In the smallest regular polygon where all user points are located, establish the origin coordinate of its center position (x_0, y_0).

(2) Divide the regular polygon into 360° according to polar coordinates and divide it into k parts.

(3) Calculate the angle between the ray and the polar coordinate axis from the origin coordinate to each data point, and then classify according to the angle.

$$\theta = arcos\frac{|y_i - y_0|}{\sqrt{(x_i - x_0)^2 + (y_{(i)} - y_0)^2}} \tag{8}$$

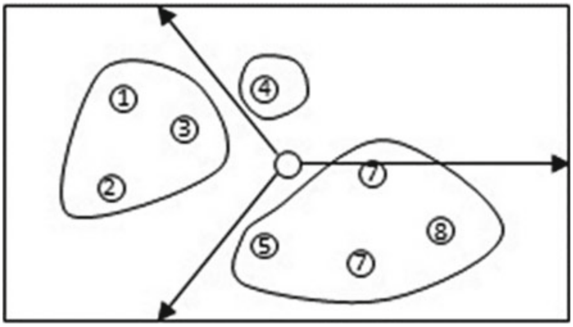
$$\alpha = \begin{cases} \theta & g_i \in first\ quadrant \\ 180^{\circ} - \theta & g_i \in Second\ quadrant \\ 180^{\circ} + \theta & g_i \in Third\ quadrant \\ 360^{\circ} - \theta & g_i \in Fourth\ quadrant \end{cases} \tag{9}$$

$$\mathrm{n} = \begin{pmatrix} 1 & \alpha \in \left(0^{\circ}, \frac{360^{\circ}}{k}\right) \\ 2 & \alpha \in \left(\frac{360^{\circ}}{k}, 2 * \frac{360^{\circ}}{k}\right) \\ \cdots & amp; \\ k & \alpha \in \left(360^{\circ} - \frac{360^{\circ}}{k}, 360^{\circ}\right) \end{pmatrix} \tag{10}$$

(4) Use the center of gravity method to obtain the initial cluster center coordinates of each class for the user points in each class
As shown in Fig. 2, assuming that it is divided into three categories, the center position in the area is first established, and then the area is divided into three parts at every 120°, and the user point determines the angle between the origin line and the polar coordinate axis, And then determine the initial category, and then find the initial cluster center coordinates in each category according to the center of gravity method.

In addition, the k-means clustering algorithm only considers the coordinate factors of each user point when calculating the clustering center, but in the actual terminal distribution situation, each distribution point has a different demand, and some people may appear The demand for distribution points in dense areas is high and the demand

Fig. 2 Gravity method to find initial cluster center



for certain personnel scattered areas is low. In this case, if only the coordinate factors are considered to find the location of the distribution center, there will be a situation where the distribution center is relatively far away from the cluster of user points with high demand, resulting in vehicles needing to take more paths during later delivery, resulting in More expenses. Therefore, considering the actual situation, the formula for selecting the location of the distribution center is improved as follows:

$$x_j = \sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)} w^{(i)} \Big/ \sum_{i=1}^{m} 1\{c^{(i)} = j\} w^{(i)} \tag{11}$$

$$y_j = \sum_{i=1}^{m} 1\{c^{(i)} = j\} y^{(i)} w^{(i)} \Big/ \sum_{i=1}^{m} 1\{c^{(i)} = j\} w^{(i)} \tag{12}$$

where w_j is the amount of express delivery required by each user.

B. *Improvement of ant colony algorithm*

Highways can be divided into five grades: national highways, provincial highways, county highways and township highways, and special highways according to administrative levels. The average speed of vehicles traveling on different road levels is different. Then the improved η_{ij} is as follows:

$$\eta_{ij} = \frac{v_{ij}}{d_{ij}} \tag{13}$$

where the d_{ij} is the distance between city i and city j, and the v_{ij} is the average speed between city i and city j.

In addition, the ant colony algorithm only considers the coordinate factor of each user point and the distance factor between two points when seeking the optimal path, but in the actual terminal distribution situation, each distribution point has a different demand, which may be There are situations where the demand for distribution points in some densely populated areas is high and the demand in certain scattered areas is low. In this case, the demand of each user point has a certain influence on the choice of distribution path, and the demand level of the user points with high demand is higher. Therefore, considering the actual situation, the ant probability formula of the ant colony algorithm is improved as follows:

$$p_{ij}^{k}(t) = \left\{ \begin{array}{cc} \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta} w_j}{\sum\limits_{s \in J_k(i)} [\tau_{is}(t)]^{\alpha}[\eta_{is}]^{\beta} w_j}, & j \in J_k(i) \\ 0 & \text{else} \end{array} \right\} \tag{14}$$

where w_j is the amount of express delivery required by each user.

4 Simulation Experiment and Result Analysis

In order to verify the correctness and feasibility of the improved k-means clustering algorithm and improved ant colony algorithm proposed in this paper, the distribution center addressing environment and vehicle route planning environment were constructed using MATLAB software, the traditional k-means clustering algorithm and improved k -means clustering algorithm is used to solve the distribution center location experiment in the region. Traditional ant colony algorithm and improved ant colony algorithm for solving vehicle path planning experiment.

A. *Distribution center location environment*

In the location environment of the distribution center, there are 40 terminal outlets in a certain area, and it is now required to integrate and establish three terminal distribution centers to be responsible for the distribution of the 60 terminal outlets. The terminal outlets are represented by circles, and the distribution centers are represented by asterisks. The newly built terminal distribution center can meet the needs of each outlet and the cost is only related to the distance. The results of the traditional k-means clustering algorithm and the improved clustering algorithm are shown in Figs. 3 and 4.

According to the simulation results, it can be seen that the coordinate results of the three distribution centers obtained from the location selection of the traditional k median center are: (116.3427, 39.7348), (116.3925, 39.7533) and (116.3120, 39.6699), and the number of iterations is three. The corresponding user points of the

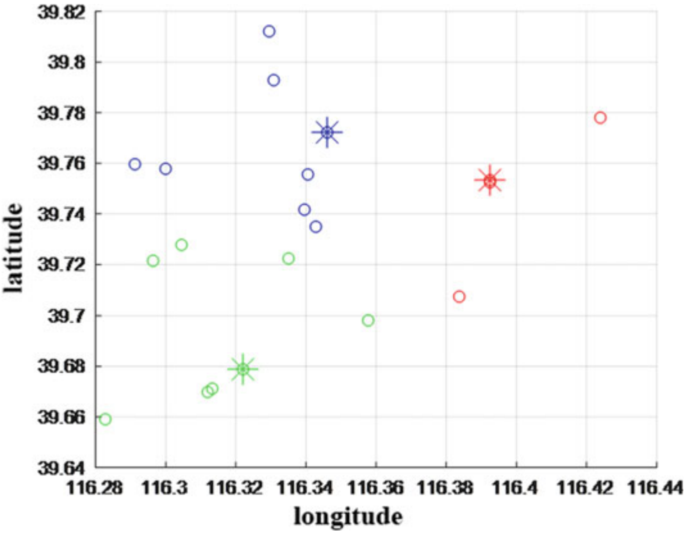


Fig. 3 Location result of traditional k-means clustering algorithm

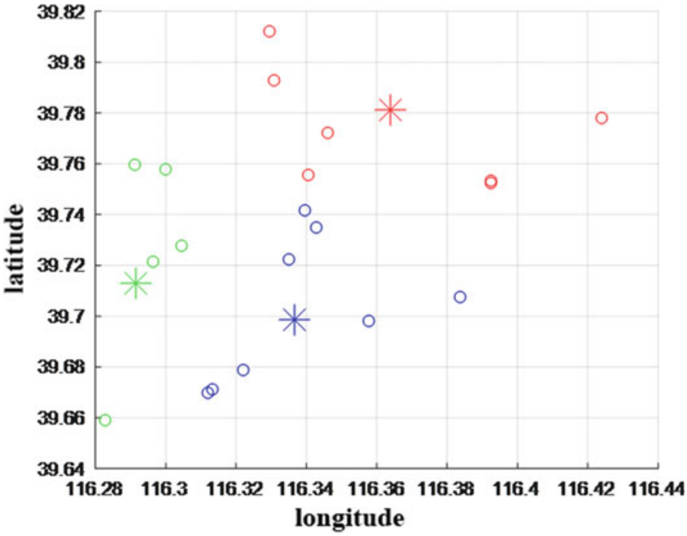


Fig. 4 Improved K-means clustering algorithm results

three distribution centers are: (5, 7, 13, 10), (3, 8, 9, 10, 12, 14, 18, 19) and (1, 2, 4, 6, 11, 15, 16, 17).

The improved k-means clustering algorithm obtains three result coordinates: (116.3638, 39.7809), (116.2915, 39.7131), (116.3368, 39.6984), and the number of iterations is twice. The corresponding user points of the three distribution centers are: (4, 5, 6, 7, 15, 17, 20), (1, 3, 11, 12, 19), (2, 8, 9, 10, 13, 14, 16, 18).

B. *Vehicle route planning environment*

In the vehicle distribution route environment, there is a logistics distribution center located at (70, 40). Now we need to use the vehicle to deliver products to 11 customers. Suppose each car has a weight limit of 1t. The circle indicates the distribution center and outlets. The blue line indicates the distribution route. 0 indicates the distribution center. Vehicles must depart from the distribution center and return to the distribution center. The simulation results are shown in Figs. 5 and 6.

It can be seen from the simulation results that the results obtained by the traditional ant colony algorithm are divided into three lines. The first line includes user points {3, 8, 9, 5}, and the second line includes user points {11, 6, 1, 2, 4}, the third line includes user points {10, 7}. The improved ant colony algorithm also obtains three results: the included user points are {3, 8, 9, 5}, although the second line includes the same user points as {1, 2, 4, 6, 11}, but The order of passing the user points is different, it is {6, 11, 1, 2, 4}. The third line has the same result as the basic ant colony algorithm, which is {10, 7}. The improved ant colony algorithm takes into account the driving speed of the vehicle and the needs of user points when solving the path, so first pass user point 6 and then user point 11.

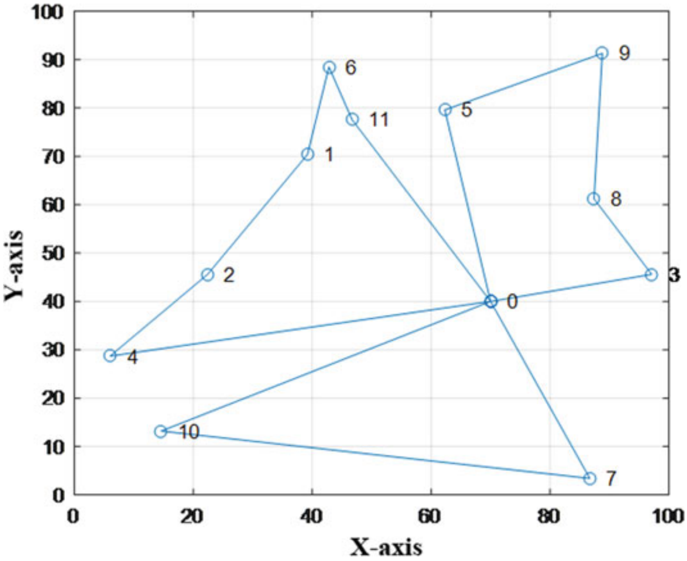


Fig. 5 Traditional ant colony algorithm

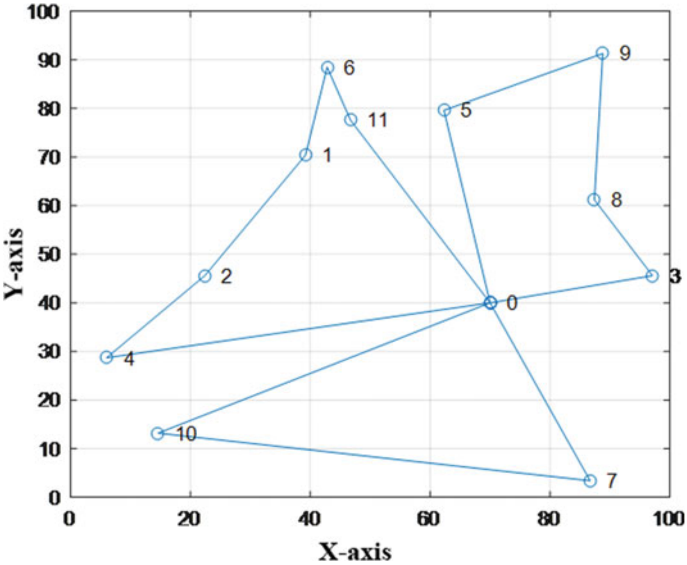


Fig. 6 Improved ant colony algorithm

5 Conclusions

In order to improve the efficiency of distribution at the end of the logistics, an improved k-means clustering algorithm is proposed for the location of the distribution center, and an improved ant colony algorithm is proposed for the problem of vehicle path planning. In order to solve the problem of random location of initial cluster centers of traditional k-means clustering algorithm, the center of gravity method is introduced to improve the original initial cluster center location, so that the selection of initial cluster centers is no longer random. At the same time, in order to make the site selection more adaptable to the logistics end distribution link, the demand of user points is introduced to improve the cluster center location function. In order to solve the problem that the ant only selects the next target point according to the distance between two points in the ant colony algorithm, the vehicle travel speed is introduced to improve the expected degree of the ant. At the same time, in order to be able to adapt to the logistics end distribution link in the vehicle path, introducing the demand of the user point improves the probability function of the ant to select the next target point. Finally, MATLAB simulation experiment was carried out. The experimental results show that the improved algorithm can more effectively and sensitively select the location of the distribution center and the vehicle travel path in the location of the distribution center and vehicle path planning, which proves the effectiveness and feasibility of the algorithm.

Acknowledgements This work was supported by the Scientific Research Program of Beijing Education Commission under Grant KM201910015004.

References

1. Y.W. Zhang, Research on location selection of distribution center based on supply chain environment. M.S. thesis, Yunnan University Of Finance And Economics (2011)
2. D.L. Chen, Research on location selection of express delivery center in the same city based on CFLP model. M.S. thesis, Chongqing Jiaotong University (2011)
3. D. Yang, Solving location problem of logistics distribution center based on improved gravity search algorithm. M.S. thesis, Shenyang University (2013)
4. W.Q. Gao, Research on the dynamic construction process of regional logistics distribution system. M.S. thesis, Henan University (2017)
5. X.Q. Huang, Research on the optimization method of city logistics distribution multi-center location. M.S. thesis, Chongqing Jiaotong University (2019)
6. X.X. Bao, X.C. Xing, Evaluation of green logistics system of solid waste at ports based on analytic hierarchy process. Environ. Eng. Manag. J. **18**(11), 2491–2499 (2019)
7. G. Dantzig, J. Ramser, The truck dispatching problem. Manag. Sci., p. 80–91 (1959)
8. J.L. Zhang, Y.W. Zhao, H.Y. Wang, Q. Jie, W.L. Wang, Modeling and optimization of vehicle routing problem with multi-model dynamic demand. Comput. Integr. Manuf. Syst. **16**(03), 543–550 (2010)
9. J.H. Ma, Y. Fang, J. Yuan, Mutation ant colony algorithm for the fastest vehicle path problem in multi-lane and multi-vehicle. Syst. Eng.-Theory Pract. **31**(08), 1508–1516 (2011)

10. S.S. Liu, Research on vehicle routing problem in limited gas station network. M.S. thesis, Tsinghua University (2013)
11. Y. Tang, Research on optimization of urban distribution route considering traffic control. M.S. thesis, Dalian Maritime University (2015)
12. Y.Z. Zhang, Research on ant colony optimization algorithm for vehicle routing problem based on spark with time window constraint. M.S. thesis, Xiamen University (2017)
13. Y. Guo, Research on vehicle routing problem of the last kilometer in rural E-commerce logistics. M.S. thesis, Beijing Jiaotong University (2017)
14. W. Hu, An improved flower pollination algorithm for optimization of intelligent logistics distribution center. Adv. Prod. Eng. Manag. **14**(2), 177–188 (2019)
15. G.I. Fragapane, C. Zhang, F. Sgarbossa, J.O. Strandhagen, An agent-based simulation approach to model hospital logistics. Int. J. Simul. Model. **18**(4), 654–665 (2019)

Customer Churn Prediction in the Broadband Service on Machine Learning

Yujie Guo and Lei Huang

Abstract With the development of the telecom industry, the telecom industry is gradually saturated. In order to increase profits, telecom companies must reduce the churn of old customers. As an important part of telecom customers, the study of broadband customer churn prediction is helpful for enterprises to find the customer churn and make effective measures timely. In this study, two feature classification methods, RF-RFE and SVM-RFE, and four prediction classification algorithms, namely Logistic Regression, KNN, SVM model and XGBoost model were compared. GridCV is used to adjust the parameters of the method to achieve the optimal effect. The assessment found that the XGBoost model performed better than other models in predicting broadband customer churn. This provides a good reference for predicting the churn of broadband customers.

Keywords Broadband customer churn prediction · Machine learning · Feature selection · Classification · XGBoost model

1 Introduction

The most important thing for telecom enterprises is the number of customers. The more customers telecom companies have, the more profits the telecom companies gain. Because the telecom industry is gradually saturated, the cost of developing a new customer is much higher than the cost of maintaining an old customer. Therefore, it is necessary to take effective measures to reduce the customer turnover rate if the telecom industry wants to increase profits and find the customers churn in advance. Broadband customers are an important part of the telecom industry, and the churn of customers in broadband services largely determines whether customers will continue to stay in the telecommunications enterprises. Therefore, the prediction of broadband

Y. Guo (✉) · L. Huang
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 18120610@bjtu.edu.cn

L. Huang
e-mail: lhuang@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_56

customers churn is helpful to retain the old customers of the enterprise, improve the loyalty of customers, discover timely the deficiencies of the enterprise according to the research situation, make timely improvements, improve the competitiveness of the enterprise, and reduce the loss of the enterprise.

The method of machine learning is suitable for analyzing and predicting the problem of customer churn and discovering customer information of other service industries [1]. Bingquan studied the problem of customer churn in land-line communication service field by using the method model of decision tree, multi-layer perceptron neural network and support vector machine [2]. Buckinx used Logistic Regression and random forest models to analyze and predict the problem of customer churn in FMCG retail enterprises [3]. Gordini developed a customized customer churn prediction model for B2B e-commerce industry using SVM model [4].

For telecom companies, customer churn prediction and management are crucial in the widely open mobile communication market, and they need to keep valuable customers in the fierce market summary [5]. Machine learning technology can be developed for other telecommunications applications. Meng used Logistic Regression to predict the customer churn of a commercial bank [6]. Lu N studied customer churn prediction and proposed a method of boosting to enhance the customer churn prediction model, and achieved good results [7].

In this paper, by comparing two different feature selection methods of RF-RFE and SVM-RFE, Logistic Regression model, KNN, SVM and XGBoost model. The rest of this article is arranged as follows. In the next section, the relevant models are introduced. In the following sections, different feature selection methods and four prediction classification models are used to predict the specific data set of broadband customer churn, and comparative evaluation is conducted. Finally, this paper summarizes the conclusions, and put forward the shortcomings of this paper and the future research direction.

2 Methodolog

A. *SVM-RFE*

Guyon takes the classification performance of support vector machines as the evaluation criterion for feature selection, and proposes a backward recursive feature selection elimination algorithm SVM-RFE [8]. SVM-RFE is essentially an encapsulation pattern selection method for heuristic search strategies, which uses recursive elimination to remove features one by one as a classifier.

SVM-RFE is a supervised sequence backward selection algorithm. In the linear classifier, it takes the discriminant information of each feature to the objective function as the sorting coefficient. $y = \mathrm{w} \cdot x + \mathrm{b}$ w is the weight vector and y is the classification surface. The feature sorting table is constructed according to the contribution of the weight vector to the classification surface. If the corresponding weight of the feature is larger, the decision function will be more influenced, and the feature with larger weight will have more discriminant information. Each iteration removes a feature with the minimum weight of $||w||^2$, and retrains the classifier until the feature sorting table is completed [9].

Guyon assumes that in the training sample matrix, when a feature is removed, the median value of the quadratic programming remains unchanged, the obtained classifier does not change. Under the premise of this assumption, the contribution value of each feature to the objective function is the sorting coefficient.

$$\begin{cases} Rank(i) = \frac{1}{2}a^T Qa - \frac{1}{2}a^T Q(-i)a \\ Q_{ij} = K\left(x_i, x_j\right) = \Phi(x_i)^T \Phi(x_j) \end{cases} \tag{1}$$

This hypothesis is also reasonable and feasible in practical application. In the formula, $\mathrm{a} = [\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{i}]$; If the i feature is removed, the meaning of $Q(-i)$ is the value of the Q matrix calculated. It should be noted that a single feature does not necessarily make the SVM classifier get the best classification performance for the features listed above, but a combination of features makes the classifier get the best classification performance. Therefore, the SVM-RFE algorithm can select complementary feature combinations [10].

B. *RF-RFE*

Random forest, a classification method proposed by Breiman L in 2001, is a classification model containing multiple decision trees [11].

(1) *Select n samples from the sample set by Bootstrap sampling*;
(2) *K variables are randomly selected from all attributes and a decision tree is constructed with these k variables and n samples*;
(3) *Repeat the previous two steps until the m decision tree is built*;
(4) *A decision tree of m was used for classification prediction, and the final prediction results were obtained by weighting or voting for m results.*

The feature recursive elimination method (RF-RFE) based on the random forest algorithm initializes the required feature set into the entire data set, and removes the data with the smallest sorting criterion score each time until the final feature set is obtained [12]. In RF-RFE, the sorting criterion score of the i features is defined as: $c_i = w_i^2$. The features with the smallest sorting criterion score are removed in each iteration, and then the remaining features are used to train RF for the next iteration.

C. *KNN*

KNN (k-nearest-neighbor) machine learning method is commonly known as k-nearest Neighbor algorithm. Cover and Hart first proposed that the time consumed in the learning stage is 0, and the prediction and recognition will be carried out after receiving the sample to be tested with the unknown tag [13].

The samples of N known tags are respectively classified into class W_i, $i = 1, 2, \ldots$, n. In the new sample, K samples closest to each other are selected. Assuming that K_i belongs to W_i class, the decision mode of W_i class is $g_i(x) = K_i,\ i = 1, 2, \ldots, n$.

The decision rule is: if $g_k(x) =_{i=1,2,3,\ldots,n}^{\max} g_i(x)$, then $x \in w_k$ [14].

D. *SVM*

Vapnik proposed support vector machines in 1997 [15]. This is a linear and nonlinear classification method. The basic idea is to map the data to be classified to a higher dimensional feature space with certain fault-tolerant conditions by using an appropriate kernel function, and to classify the data by constructing an optimal classification hyperplane in this space.

Its goal is to find a hyperplane, we need to find the maximum support vector to the separated hyperplane distance, under this condition to find the separated hyperplane. The optimal hyperplane is $w^T x + b$, and the values of w and b are required.

$$\min \frac{1}{2}||w||^2 \quad y_i(w^T x_i + b) \geq 1, \; i = 1, \ldots, n \tag{2}$$

The constraint conditions for solving the maximum geometric interval, Formula (2) are added to the objective function, and the Lagrange formula is established:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1) \quad s.t.\alpha_i \geq 0 \tag{3}$$

Because the goal meet the KKT conditions, the optimization goal $\underset{w, b}{\min}\left(\underset{\alpha_i \geq 0}{\max} L(w, b, \alpha)\right)$ into the dual problem $\underset{\alpha_i \geq 0}{\max}\left(\underset{w, b}{\min} L(w, b, \alpha)\right)$ of w, b for partial derivatives, the value is 0, restore the w, b in $L(w, b, \alpha)$, finally to α_i derivation $W(\alpha)$ is of the maximum.

E. *XGBoost*

Chen improved the GBDT (Gradient Boosting Decision Tree) algorithm and proposed an efficient, flexible and portable optimal distributed Decision Gradient promotion library (XGBoost) [16, 17].

XGBoost algorithm's basic idea is a given training set assigned to each input sample different leaf node split points according to the value of an attribute. Each leaf node corresponds to a real-time score, when given to predict sample x_i, the result is the sum of forecast score of each tree for the predictive results. The specific model can be defined as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \; f_k \in F \tag{4}$$

F is all space for classification and regression tree, $\hat{y}_i$ predictive results of the corresponding x_i. $f_k(x_i)$ sample x_i this input to the first k obtained the leaf nodes of the tree after the score.

3 Data

A. *Data Source*

The data set for this paper is derived from the Kaggle database. Kaggle is a data modeling and data analysis competition platform. Companies and researchers can publish data on it. Data set has a total of 19 properties, including customer information (customer number, monthly income, etc.), contract (the contract type, network bandwidth, etc.) information and service information (whether or not to use mobile phone service, number of customer complaints, etc.). It has 17 characteristics description attribute, two target attribute (customer churn and customer because of the churn of bills).

For the factors affecting the churn, the variables are both discrete and continuous. Some variables are discrete, such as newacct_no, line_stat, bill_cycl, serv_type, churn, etc. And some variables are continuous, such as tenure, effc_strt_date, effc_end_date, complaint_cnt, etc.

B. *Data Processing*

The sample of the data set shows that there are some missing values. Properties with missing values are 'effc_strt_date', 'effc_end_date', 'contract_month', 'ce_expiry', 'term_reas_code', 'term_reas_desc'. Among them, the missing values of 'effc_strt_date', 'effc_end_date', 'contract_month' and 'ce_expiry' can be found to be the missing values of the same sample, and there are 1937 pieces of data with this part missing. This part of data belongs to the data with a small proportion, which is relative to the total data volume. The method of direct deletion is adopted for the missing value of this class.

The variable 'bill_cycl' has the value of 1. The variable 'serv_type' has a value of BBS only. The variable 'line_stat' is of current type and was not considered in this study. The 'service_code' has no specific meaning. The above variables are processed and the useless variables are deleted.

In this study, the effect of specific start time and end time of broadband customers on churn results is not considered. The 'effc_strt_date', 'effc_end_date' are not considered. Based on the characteristics of customers in the current month, this paper makes a judgment on whether churn or not. It does not consider the current time variable.

The bandwidth of 100 M is 69% of the total, so the 100 M is classified as 0 and the rest as 1. In the data set, the variables 'term_reas_desc' and 'term_reas_code' have the same meaning. Only those with contract termination can have 'term_reas_code' and 'term_reas_desc'. Therefore, the fixed value is used to supplement the missing value, and the 'noCode' is used to indicate that the contract is not terminated for such missing cases. According to the contract termination description, the codes are divided into 5 categories. The samples that have not terminated the contract are classified as 'noCode', the ones related to the customer's reason are classified as 'customerCode', including 'CUCO' ('Downsizing/Cut cost'), 'NU' ('No Use'), 'OT' ('Overdue Termination: Involuntary termination by credit control'), 'NCAP' ('No capacity'), 'CUSN2' ('Customer Issue: No use'), 'CUSB0' ('Customer Issue: Bankruptcy'), 'TRM' ('Termination'), 'CLB' ('Closing Business'), and the ones related to the sales' reason are classified as 'saleCode', including 'EXP' ('Sales Plan Not Attractive'), 'COVL3' ('Coverage Issue: Low speed coverage (Customer Requests 30 M)'), 'COM15' ('Com-Unsatisfy Service quality'), 'COVL2' ('Coverage Issue: Low speed coverage (Customer Requests 100 M)'), 'COVL1' ('Coverage Issue: Low speed coverage (customer requests 200 M+)'), 'COM10' ('Com-Miss follow-Help Desk') service's reasons related to be classed as 'serviceCode', contains the 'NET' ('Network Problem'), 'UFSS' ('Unsatisfy Field Service Support'), 'UCSH' ('Unsatisfy Cs Hotline'), 'LOSF' ('Lack of Service Features'), 'PLR' ('Parallel Run Order'), 'UEMS' ('Unsatisfy Email Service'), 'NWQU' ('Network quality'), other causes to be classed as 'otherCode' contains 'OTHS', 'BILP' ('Billing Problem'), 'the EXI' ('Additional extra installation charge (Part II)'), 'MGR' ('Migration Order'), 'REV' ('Relocate to non-coverage(w/prof)').

Since there are two columns in the data set to identify whether the customer churn or not, the variables of the two columns are processed. The marked churn is 1, and the non-churn value is 0. The sum of the two variables greater than or equal to 1 is the final churn sample, and the value of 0 is the non-churn sample.

There are dimensional differences. For binary variables like 'with_phone_service', the method of OneHotEncoder is used. Because the dimensions of continuous variables are different, continuous variables need to be standardized in order to accelerate the speed of finding the optimal solution of gradient descent. The dataset standardizes the features by removing the average and scaling to the unit variance.

In this paper, 70% of the samples in the data set are selected as the training set, and the remaining 30% are the test set. The sample of this data set is unbalanced between the segments of broadband customer churn and not churn. It is a highly unbalanced data, among which the sample of non-churn customers is far more than that of churn customers. In order not to make the final prediction results excessively biased to the non-churn samples, the data of the training set is balanced by using the SMOTE oversampling algorithm. After oversampling, the proportions of classes 0 and 1 are 50 and 50%.

This study adopts the method of wrapped feature selection, which is a common method in feature selection. The wrapped method takes the optimization criteria of the model as the criteria for feature selection, and selects several features at a time or excludes several features.

This paper uses recursive feature elimination. In order to compare the differences of different feature selection methods and prediction results of different models, RF-RFE and SVM-RFE were used for multiple rounds of training. After each round of training, the features of several weight coefficients were eliminated. The next round of training was conducted based on the new feature set. In order to make the feature selection more accurate, the two feature selection methods use 5-fold cross validation. The RF-RFECV and SVM-RFECV methods were used to select the features with high average scores for model training and test.

RF-RFECV method was used to select 10 characteristics. 'bandwidth', 'tenure', 'secured_revenue', 'contract_month', 'with_phone_service', ‘term_reas_code_customerCode’, ‘term_reas_code_noCode’, “term_reas_code_otherCode’, ‘term_reas_code_serviceCode’, ‘ce_expiry’.

The ‘term_reas_code_saleCode’ and ‘complaint_cnt’ are eliminated, the RF-RFECV screen out the class features low weight, which has little influence on the subsequent model training.

SVM-RFECV method was used for feature selection, the weight coefficients of each feature were obtained. And the features were selected from large to small according to the coefficients. Eight characteristics was selected. The characteristics of ‘term_reas_code_otherCode’, ‘contract_month’, ‘term_reas_code_customerCode’, ‘complaint_cnt’ was eliminated. According to the SVM-RFECV method, the influence of these factors is less than that of other features, and the selected features are also used to train, test and evaluate the four models.

4 Results

A. *RF-RFECV*

This paper modeled “churn” with these selected 10 characteristics on Logistic Regression, KNN model, the SVM model and XGBoost model.

The parameters in each model were adjusted by grid search method to optimize the performance, and the evaluation model was evaluated by ROC curve. Comparison of ROC curves for each model is shown in Fig. 1.

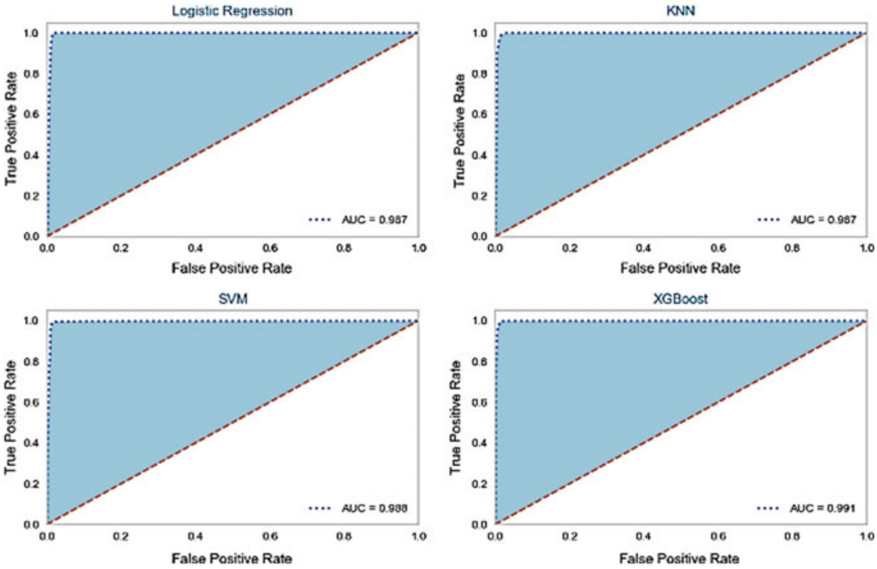


Fig. 1 Roc curves of the four models

Table 1 AUC, accuracy, precision, recall, and F1-score of the four models

Model	AUC	Accuracy	Precision	Recall	F1-score
Logistic regression	0.987	0.986	0.972	0.987	0.980
KNN	0.987	0.984	0.968	0.987	0.977
SVM	0.988	0.985	0.971	0.988	0.979
XGBoost	0.991	0.989	0.976	0.991	0.983

As can be seen from the ROC curve, the performance of the four models is very good, and all of them are far greater than 0.5, among which the XGBoost model has the best classification effect among the four models, and the AUC score is over 0.99.

The evaluation values of the model are shown in Table 1. The scores of the four models were not significantly different in recall and f1-score. The scores of AUC, accuracy and precision were relatively high in XGBoost model. Logistic Regression, KNN and SVM model have a small difference in the accuracy of the predicted samples. This suggests that the four models are relatively equal in their ability to predict the positive samples.

In order to be able to better contrast between broadband customer churn prediction model, this paper lists Table 2 that describes the classification or confusion matrix for the four models in identifying been whether churn or not.

Through Table 2, we can find XGBoost model in the prediction of 'non-churn' and 'churn' performance is better than other model performance, which has good prediction effect in the 19,149 test sample. Each model in the prediction of performance 'non-churn' samples are in good shape. But XGBoost has an even greater advantage in predicting customer churn. Losing customers is what the broadband industry will focus on.

B. *SVM-RFECV*

Grid search is used to find the optimal parameters, and the evaluation value is used to evaluate each model. The ROC curve comparison of each model

Table 2 Classification matrix for the four models

Model	Observed	Predicted	
		Non-churn	*Churn*
Logistic regression	Non-churn	14,864	222
	Churn	42	4021
KNN	Non-churn	14,826	260
	Churn	37	4026
SVM	Non-churn	14,849	237
	Churn	34	4029
0XGBoost	Non-churn	14,887	199
	Churn	21	4042

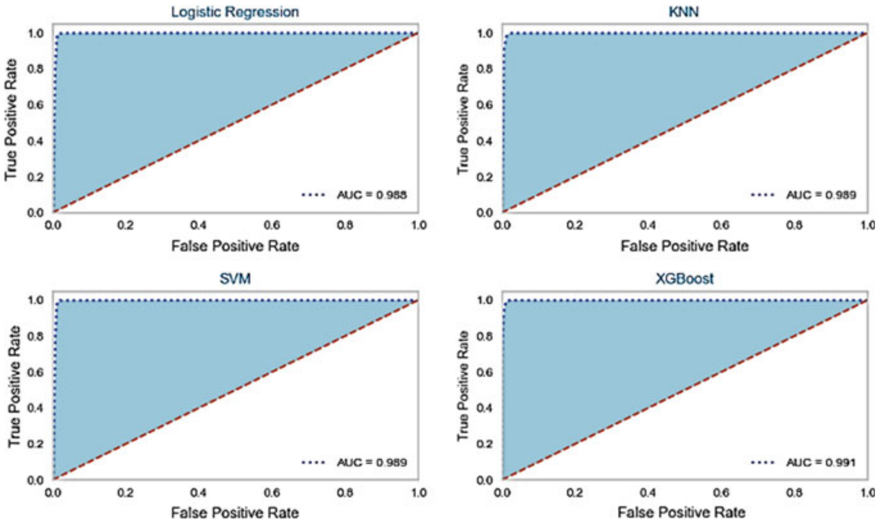


Fig. 2 Roc curves of the four models

is shown in Fig. 2. Compared with RF-RFECV method for feature selection, Logistic Regression, KNN and SVM methods all improved their AUC scores to some extent, while XGBoost model did not significantly improve, but the AUC score of XGBoost model was still the highest among the four models.

Table 3 shows the other evaluation values of the four models. KNN, SVM and XGBoost model decreased in accuracy value compared with RF-RFECV method. The SVM and XGBoost model also decreased in precision value.

Through Table 4, we can see the reason for the decline.

We can see four models in the prediction of ‘churn’ class accuracy is promoted. The accuracy of the SVM model and XGBoost model fell in predicting ‘non-churn’ class. Since Table 3 values are the average of the two categories, the SVM model and XGBoost model are reduced in precision.

Two different feature selection methods, RF-RFECV and SVM-RFECV, were compared comprehensively. And four models were trained, fitted and tested for different features. The XGBoost model performed best of the four, but the two different feature selection methods did little to improve the XGBoost model. The

Table 3 AUC, accuracy, precision, recall, and F1-score of the four models

Model	AUC	Accuracy	Precision	Recall	F1-score
Logistic regression	0.988	0.986	0.973	0.988	0.980
KNN	0.989	0.986	0.971	0.989	0.980
SVM	0.989	0.986	0.969	0.989	0.979
XGBoost	0.991	0.988	0.975	0.991	0.983

Table 4 Classification matrix for the four models

Model	Observed	Predicted	
		Non-churn	*Churn*
Logistic regression	Non-churn	14,865	221
	Churn	40	4023
KNN	Non-churn	14,850	236
	Churn	29	4034
SVM	Non-churn	14,829	257
	Churn	20	4043
XGBoost	Non-churn	14,881	205
	Churn	20	4043

XGBoost model can still be used as a priority algorithm for predicting whether broadband customers churn or not.

5 Conclusion

In this paper, different feature selection methods and different training models are used to predict broadband customer churn. The evaluation scores of the models under various conditions were compared. In order to fully compare the strengths and weaknesses of RF-RFECV and SVM-RFECV for Logistic Regression, KNN, SVM model and XGBoost model, the accuracy, precision, recall, f1-score and roc curves were used to evaluate the performance of the models.

At the same time the paper mapped the four models of Classification matrix, detailed comparison on the numerical model in the 'churn' and 'non-churn' performance. All four models performed well in this data set. The XGBoost model performed better than the others in both cases. In practical application, enterprises can give priority to the XGBoost model to predict the churn of broadband customers, which can help to find out the churn of customers in time, and take effective retention measures for this part of customers. According to the selected characteristics, enterprises can take certain preferential measures in advance for these factors with greater impact to prevent customer churn.

The research of this paper has some shortcomings. On the one hand, it can be considered the connection between the churn of broadband customers and the time period, and it can use the time-series approach to analyze the high potential churn of broadband customers in a short period of time. On the other hand, the method of improving feature selection is considered to reduce the time of feature selection in the case of large amount of data. The other method of feature selection is considered to greatly improve the accuracy of the existing algorithm and reduce the training time of the algorithm.

References

1. S.S. Anand, A.R. Patricj, J.G. Hughes, D.A. Bell, A data mining methodology for cross-sales. Knowl.-Based Syst. **10**, 449–461 (1998)
2. B. Huang, T. Kechadi, B. Buckley et al., A new feature set with new window techniques for customer churn prediction in land-line telecommunications. Expert Syst. Appl. **37**(5), 3657–3665 (2010)
3. W. Buckinx, D. Van den Poel, Customer base analysis: partial defection of behaviourally loyal clients in a noncontractual FMCG retail setting. Eur. J. Oper. Res. **164**(1), 252 268 (2005)
4. N. Gordini, V. Veglio, Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind. Market. Manag. **62**, 100–107 (2017)
5. F. Napitu, M.A. Bijaksana, A. Trisetyarso, Y. Heryadi, Twitter opinion mining predicts broadband internet's customer churn rate, in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Phuket, pp. 141–146 (2017)
6. X. Meng, S. Cai, K. Du et al., Research on the prediction model of customer churn for industrial banks. Syst. Eng. **22**(12), 67–71 (2004)
7. N. Lu, H. Lin, J. Lu et al., A customer Churn prediction model in telecom industry using boosting. IEEE Trans. Ind. Inf. **10**(2), 1659–1665 (2014)
8. I. Guyon, J. Weston, S. Barnhill et al., Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
9. L. Xu, Research on Feature Selection Algorithm Based on SVM-RFE and Particle Swarm Optimization (Hunan Normal University, 2014)
10. Y. Lu, Application of SVM-RFE Algorithm in Data Analysis (Jilin University, 2009)
11. Breiman L. Random forests. Mach. Learn. **45**(1) (2001)
12. C. Wu, J. Liang, W. Wang, C. Li, Random forest algorithm based on recursive feature elimination method. Stat. Decis. **21**, 60–63 (2017)
13. L. Zhao, X. Yu, S. Zhang et al., Research on SVM-KNN classification algorithm. Comput. Dig. Eng. **248**(6), 31–39 (2010)
14. J. Li, S.X. Pan, L. Huang, X. Zhu, A machine learning based method for customer behavior prediction. Tehnicki vjesnik-Technical Gazette **26**(6), 1670–1676 (2019)
15. V.N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 2000). https://doi.org/10.1007/978-1-4757-3264-1
16. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794
17. L.L. Qin, N.W. Yu, D.H. Zhao, Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki vjesnik-Technical Gazette **25**(2), 528–535 (2018)

Research on the Identification and Prediction of Default Risk of Online Lending Platform Customers

ShuaiQi Liu and Sen Wu

Abstract P2P lending is a new market-oriented and networked financial service platform, which has attracted the attention of researchers in different fields since its establishment. The P2P lending platforms have operation mechanism design and fund management issues. Due to the problem of information asymmetry in the credit information of the P2P platform and the lack of a physical mortgage mechanism, the user's default identification and the risk management and control of the platform have become the core problems of P2P operation, and data mining has become the main method to study this problem. This paper studies the default problem of P2P platform customers, uses the random forest method to establish a classification model on Lending Club data to predict the P2P platform customer default status, and analyzes the importance of the attributes that affect whether or not to default. The research shows that the random forest model can effectively predict the customer's default behavior, and the repaid amount has become an important factor for default.

Keywords The random forest · Attribute importance · P2P online lending · Default prediction

1 Introduction

P2P lending is the abbreviation of peer to peer lending, which is a direct lending behavior from individual to individual through the network platform. P2P is a microfinance method that skips intermediaries through the network platform, gathers funds from fund providers through the network, and lends funds to those who need it. Compared with traditional banks, P2P lending can make it easier for borrowers to obtain loans, enable investors to obtain higher interest rates, and reduce transaction

S. Liu (✉) · S. Wu
School of Economics and Management, University of Science and Technology Beijing, Beijing, China
e-mail: liushuaiqi_ustb@163.com

S. Wu
e-mail: wusen@manage.ustb.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_57

costs for both supply and demand. This way of borrowing eliminates the complicated procedures of traditional financial services, and provides convenient and fast borrowing channels with low entry barriers. P2P lending is a new business model that promotes inclusive finance and improves the efficiency of capital use. Successful P2P platforms aggregate providers and users of funds in different ways and earn intermediary fees by providing services. This model is different from the traditional banking model, which can activate the possibility of combining capital flow and commodities. The growth of P2P platforms has also brought strong competitive pressure to traditional financial service institutions, so many financial service institutions have also actively participated in the construction of this type of emerging platform.

After the establishment of the world's first P2P lending platform Zopa in London, England in 2005, the wave of P2P lending platforms in Europe and the United States began to spread. Lending Club, which was established in 2006, has become the largest P2P lending platform in the United States. The rise of P2P lending platforms is not only because of the simple borrowing process but more important is that many borrowers with poor financial quality are still unable to borrow money from banks even if they are willing to pay higher interest rates. The traditional solution is that borrowers generally seek loans from private lending, underground banks, and mutual aid associations. However, under the tide of modern Fintech, P2P lending platforms came into being, which solved the problem of such credit rationing in some ways.

P2P loans refer to direct loans from individuals to individuals. Investors with surplus funds borrow directly through the online platform to borrowers who need funds. Different from the traditional lending model, traditional financial intermediaries such as banks have complicated and time-consuming loan application procedures. The biggest feature of P2P platforms such as Lending Club is the automated audit system. Based on obtaining basic data of borrowers and investors, through registration, credit decisions, and credit score calculation, service and payment systems, the average number of loan approval days has been greatly shortened. The average loan approval time for a US bank is 45 days, while Lending Club only takes 7 days. The benefits of the P2P platform are not only that borrowers can obtain funds faster; investors have new investment channels, but also it removes the indirect costs of traditional banks, including credit cost, and borrowers cost. The cost of the original bank will be returned to the borrower and lender so that the borrower can borrow at a lower interest rate than the bank; investors can earn higher interest.

The Lending Club automated audit system eliminates the indirect costs of traditional financial intermediaries, but it creates the problem of Information Asymmetry. Borrowers understand their financial risks better than investors, and may mislead the platform in providing information, leading to an increase in default rates. Even if Lending Club is committed to reducing the default rate and improving the internal credit score model, the default rate is still much higher than that of traditional banks, such as Citibank's default rate is about 2%. From the historical data, the default rate of each level of Lending Club is greater than 2%, and even the best-rated A-level default rate is 3.21%. Therefore, the high default rate of the P2P platform will become the core issue of the future platform operation. This study focuses on the default problem of P2P platform users. Using the Lending Club data as a research sample, the random forest method is used to identify the user characteristics and important credit attributes of the platform default users and predict the user's credit behavior. The platform default rate provides help for P2P platforms to more effectively identify users and reduce risks.

2 Related Works

A. *Research on P2P default risk*
The operation of financial institutions focuses on the efficiency of capital use and risk prevention and control. When the financial institution processes the credit business to evaluate the credit of the applicant, the 5P principle proposed by banker Paul Hunn in 1970 is generally used, including the applicant (People), the application (Purpose), the source of repayment (Payment), and the protection of claims Five dimensions with the Perspective of Credit. However, how to implement various credit evaluation principles in the procedure depends on the method adopted by the financial institution, which emphasizes subjective experience and subjective statistical methods, which can be carefully evaluated for each applicant. This method is simple and flexible and has low evaluation cost, but lacks objective standards, is not easy to pass on experience, and is not prone to flaws if the system is not sound. It is no longer applicable today. The credit rating system is to transform the above empirical method into an institutionalized procedure. The applicant will be evaluated according to various attributes that affect credit. The procedure is more objective, but the final decision still requires subjective judgment, and the quality of decision-making will be affected. The expert system further converts the subjective subjectivity of the aforementioned experts into an automatic review system, but the expert review standards still carry a subjective component. The credit scoring system uses many consumers' past credit records to construct a scoring system for personal credits through statistical methods, find out the key elements, and judge the credit risk according to each applicant's score, as an approval or credit limit decision making. The scoring system provides an objective basis for investors, which simplifies the evaluation process, shortens the operation time, and improves the processing efficiency. The applicability of the model can be assessed over time and modifications can be made to help the bank drive the business.

Risk control is a very important job in the online credit lending platform. According to the literature, there are credit risks and operating risks in online credit lending, different methods should be used to solve different risk problems. The establishment of model architecture and decision systems can be used and referred by borrowers, investors, or even administrators. It can also predict the operation direction of the platform and borrowers and investors can distinguish the platform operation status.

For investors, investment decisions will be based on the borrower's financial information, family background, etc. as a reference. If the information provided by the borrower is inaccurate, information asymmetry will occur, leading to problems of adverse selection and moral hazard [1]. The adverse selection is that before the transaction, the information superiors deliberately hide the characteristics and induce the information disadvantaged to make wrong decisions and be victimized. Horal hazard means that the borrower does not have to bear the consequences of his actions, which leads to their actions being slightly arbitrary, which will cause investors' investment may not back. Due to the rapid rise of online credit, which exceeds the scope of the existing legal control, and has not established a unified global credit score like the traditional credit market, it has failed to establish a clear standard in real-time, such as FICO (Fair Isaac Corporation) the United States unified personal credit score. Most online credit platforms do not have globally unified credit score standards and risk control methods [2], which are prone to fraud risks. In the research on risk control, numerous solutions have been proposed, many of which are based on data mining [3], which is the most effective method for risk control.

B. *Research on the random forest*
Data classification is a common problem in data mining. The essence of the problem that needs to be solved in many fields is data classification, such as geological image classification, biological information classification, customer grade classification, etc. Classification has become one of the most important researches in the field of data processing applications. The traditional statistical data classification methods include cluster analysis and Bayesian classification algorithm. In recent years, with the expansion of various data storage capacity, more and more multi-dimensional complex data has been accumulated, and traditional data classification algorithms are not good at processing these multi-dimensional data, so scholars have begun to propose multi-dimensional Data classification algorithm.

The Decision tree is a classification algorithm widely used. It is a tree-like classifier, which selects split attributes for classification in each internal node, and each leaf node has data of the same category. When inputting the data to be classified, the decision tree is a path continuously identified from the root node to the leaf node, and the category of the leaf node of the path corresponds to the category of the sample to be classified. The decision tree is a simple and fast non-parametric classification method. In general, it has a good classification accuracy. However, when the data is complex or there is more noise, the decision tree is prone to excessive nodes or excessive configuration problems, causing classification accuracy to decrease.

Random Forest is a machine learning algorithm published by Breiman [4]. It combines the Bagging machine learning theory proposed in 1996 with the random subspace method proposed by Ho in 1998. The basic principle is that each decision tree is a weak classifier, a random forest is composed of multiple decision trees, and many weak classifiers are composed into a strong classifier. The final decision result is decided by voting. The random forest has two important parameters, one is the number m of candidate feature parameters selected when splitting at each node of a single decision tree, and the other is the number n of decision trees in the random forest. Compared with the traditional decision tree, the random forest algorithm has a stronger generalization ability and classification effect. Due to its good performance, the algorithm is widely used in practical fields such as biological information, medical research, business analysis, text exploration, semantic classification, economics and finance, and has achieved good results.

Although random forest has been applied in different fields. Kumar and Thenmozhi [5] studied the predictability of stock index movement direction, empirical results showed that the random forest method is superior to the neural network, discriminant analysis and logit model in predicting the direction of stock market movement. Montillo and Ling [6] proposed to use face image analysis to predict a person's age based on the random forest method. The study illustrated the wide range of practical applications and provided some random forest parameters. In the absence of prior models, these parameters are easier to initialize, categorize body features, and significantly reduce training time while maintaining the regression accuracy of the model. Rodriguez et al. [7] used the random forest machine learning classification method to conduct a coverage study of land classification. The research object was 14 different land types in southern Spain. Using remote sensing data monitoring, the accuracy, sensitivity, and the size and noise of the data set are used to compare classification accuracy. The results show that the random forest model overall performance is better than the single decision tree. Theofilatos et al. [8] used machine learning technology to study the impact of trading on euro/dollar exchange rate and used five supervised learning classification techniques (K-Nearest Neighbors algorithm, Naïve Bayesian Classifier, Artificial Neural Networks, Support Vector Machines, and Random Forest) for time pane movement prediction of the euro/dollar exchange rate. The study found that Random Forest's machine learning technology resulted in trading strategies that clearly outperformed all other strategies' annualized return and Sharpe ratio. Qin et al. [9] used Gradient Boosted Random Forest to study the non-linear trading patterns of stocks and constructed trading decisions with higher weighted market indicators, that is, signals of buying, selling, or holding, nine stocks and one index are used to measure the performance of trading decisions. The empirical results show that the proposed trading method generates excess returns in the buy-and-hold strategy.

C. *Research on the default of Lending Club*

Jina and Zhua [1] used 34406 samples from July 2007 to November 2011 of the Lending Club, divided the dependent variable Loan status into three categories: Defaulter, Need Attention, Well Paid. The independent variables were screened through random forests Characteristic variables, which can be derived from "loan period", "loan amount", "revolving credit interest rate", "rating given by Lending Club", "annual income", "loan purpose", "months since the last repayment", "Repayment plan", "revolving credit balance", "debt to income ratio" a total of ten parameter values. Empirical studies indicated that "loan period", "annual income", "loan amount", "debt-to-income ratio", " rating given by Lending Club", and "revolving credit interest rate" have a significant impact on default.

Emektera et al. [10] took Lending Club as the research object and 61451 samples from May 2007 to June 2012 as the sample data. The dependent variable is Loan status and is divided into three categories: Charged-off, Delay, Fully paid, the input variable selects "loan interest rate", "loan elapsed days", "rating given by Lending Club", "monthly repayment amount", "loan amount", "debt-income ratio", "repayment amount", "Remaining principal", "Repayment to debt ratio", "Monthly income ", "FICO score", "Open credit limit", "Total credit limit", "Revolving credit balance", "Revolving credit rate" "Number of joint inquiries within six months", "Arrearage Amount", "Number of arrears within two years", "Number of months after the last arrears of repayment", "Number of months after the last payment of the repayments". Empirical studies indicated that "rating given by Lending Club", "debt-to-income ratio", "FICO score", and "revolving credit rate" have a significant impact on default. Also, Byanjankar et al. [11] used the European P2P platform Bondora as the research object. The study used 16037 samples from March 2009 to February 2015 as the sample, and the target variable was the Actual Default., Default, the independent variable selects "credit decision", "first-time customer", "age", "nationality", "credit group", "loan amount", "loan interest rate", "loan age", "loan purpose", "application type", "new loan conditions", "marital status", "employment status", and "gross income". Empirical studies indicated that "whether it is a new offer" and "loan amount" has a significant impact on default.

The relevant methods to predict the possible fraud or default of the borrower, some of which are to judge the loan application or the relevant data of the borrower, such as content analysis, auxiliary data, and text mining. The content analysis method refers to quantifying and systematically analyzing the content delivered by the borrower and analyzing whether the content influences the entire transaction. Auxiliary analysis is an effective supplement to the most important data samples to improve the accuracy of data analysis. Text mining is the processing and analysis of natural language, which converts text content into readable and quantifiable data for more accurate analysis. In addition to analyzing the content of the application, many scholars also use data mining techniques to predict the default status of the borrower, usually the most classification algorithms, such as random forest, decision tree, SVM (Support Vector Machine), and so on. Random forest has become one of the important methods of credit risk research.

3 Research Object and Research Method

A. *Research object*

This study takes Lending Club as the research object, and the general LC business process is shown in Fig. 1. Since this article studies the repayment after borrowing, in all samples, FICO scores and missing values that are not related

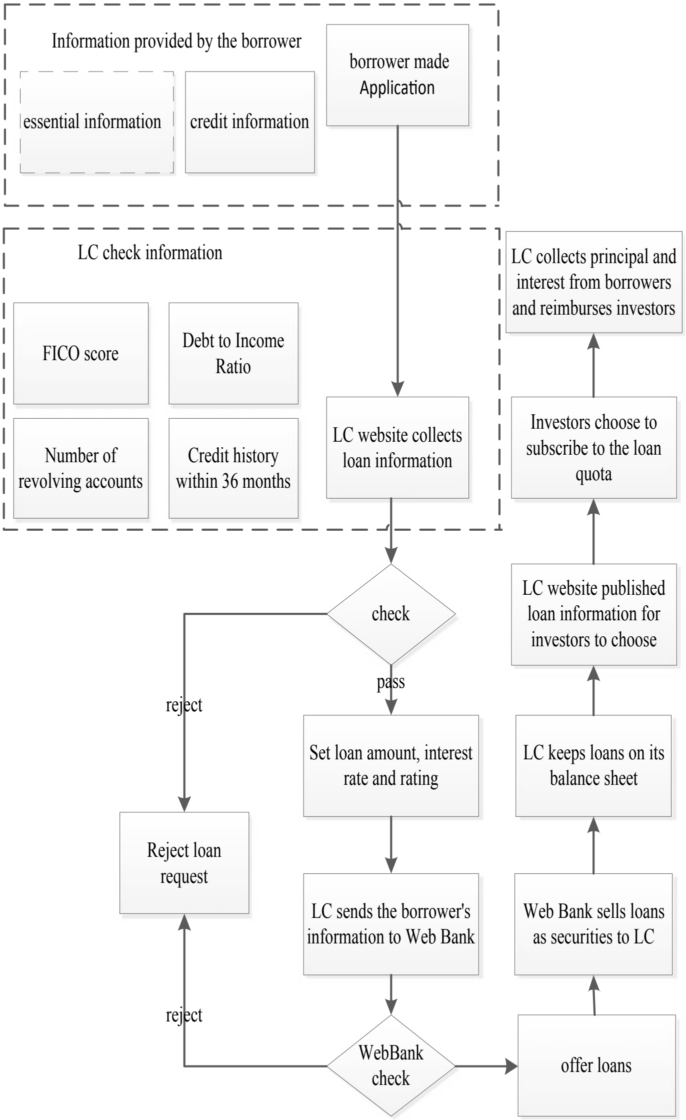

Information provided by the borrower
essential information
credit information
borrower made Application
LC check information
FICO score
Debt to Income Ratio
Number of revolving accounts
Credit history within 36 months
LC website collects loan information
check
reject
pass
Set loan amount, interest rate and rating
Reject loan request
LC sends the borrower's information to Web Bank
reject
WebBank check
offer loans
Web Bank sells loans as securities to LC
LC keeps loans on its balance sheet
LC website published loan information for investors to choose
Investors choose to subscribe to the loan quota
LC collects principal and interest from borrowers and reimburses investors

Fig. 1 General process of LC business

to the purpose of this study are deleted, selecting 44 attributes as the research object. Drawing on previous studies, we divide target variables into three categories: normal (fully paid), delayed, and default to discover the changes in the importance of the attributes throughout the repayment process.

B. *Data selection and preprocessing*

Table 1 Sample composition

	Fully paid	Delayed	Charged-off
Sample size	8412	8383	8352

To ensure the universality of the experimental conclusion, this study selected the contract data in 2019 of the Lending Club official website. The data is relatively clean, and no outliers and abnormal samples are found. Some of the attributes have a few missing values. For discrete attributes, this paper uses special values to fill in, indicating that the sample has no actual value for the attribute. For continuous attributes, this paper uses the means to fill in. Because random forest modeling has a certain sensitivity to data balance, to avoid the impact of unbalanced data and make the model have a higher accuracy, loan samples in fully paid and in default state were sampled at random, and finally, samples with the same number as delayed samples were selected to make up the data set of this paper The final sample composition is shown in Table 1.

Data preprocessing is the most basic link for subsequent analysis. Therefore, this article preprocesses the data after data collection, including data cleaning and data conversion. The data cleaning part includes attributes selection and missing value filling.

In the data transformation part, this paper converts the non-numeric data to numeric data to facilitate the subsequent processing and construction of the machine learning model. For example, the grade attribute replaces the seven values of A, B, C, D, E, F, and G with 1, 2, 3, 4, 5, 6, and 7 digital codes. For the target variable loan status in this article, the values of this attribute are Charged Off, Full Paid, Current, Default, In Grace Period, Late (16-30 days), Late (31-120 days). The values Late (16-30 days), Late (31-120 days) are the loan in the extended state, denoted by 0. The values Charged Off, Default, In Grace Period are the loan in default, denoted by 1. The value Current and Fully Paid is the loan in good condition, denoted by 2.

After converting all the sample data into numeric data, for continuous data, this article uses the z-score standardization method to normalize the data. The processed data conforms to the standard normal distribution, the mean is 0 and the standard deviation is 1.

C. *Random forest modeling and prediction*

Random forest is a model composed of many decision trees. The "forest" represents a collection of multiple decision trees. The concept of "random" has two meanings: randomly drawn samples and randomly drawn feature subsets. This article uses the randomForest package in the R for random forest modeling and the pROC package for visual presentation.

In this paper, the three target variables of fully paid state, delayed, and default are modeled in pairs. The error rates of a fully paid-delayed model, fully paid-default model, and delayed-default model are shown in Fig. 2. It can be seen from Fig. 2 that the error rate of the three models decreases with the increase of trees. When the number of trees is 200, the error rate remains basically unchanged, so the model

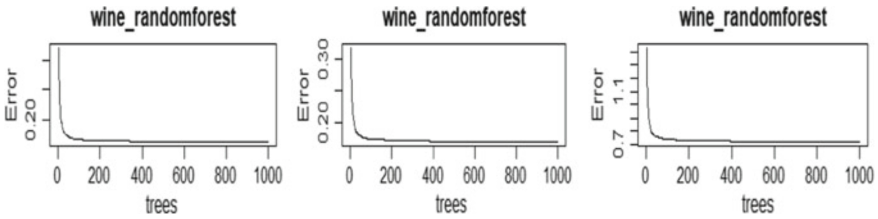


Fig. 2 Error rate under different number of trees in three experiments

ntrees parameter is 200. In the process of model parameter optimization, the change of mtry parameter value will also affect the accuracy of the model. For the random forest model, the mtry parameter generally takes the square root of the number of variables. When the ntrees parameter is determined, this paper performs different value experiments on the mtry parameter. It is found that the three models have the highest model accuracy when the mtry value is 7, so the mtry parameter value of the model is 7.

It can be seen from Figs. 3, 4, and 5 that the AUC values of the three models reached 0.754, 0.745, and 0.706, showing that the three models achieved good classification results. The confusion matrix of three experiments is shown in Table 2.

D. *Prediction of samples*

After modeling the three models, this paper selects 500 data from each of the three types of samples for model prediction, regularizes the attribute values of the predicted samples into a form consistent with the training samples, and then replaces the test set in the modeling process with predicted samples, and then the model is rerun to predict the sample. The final prediction result is shown in Table 3.

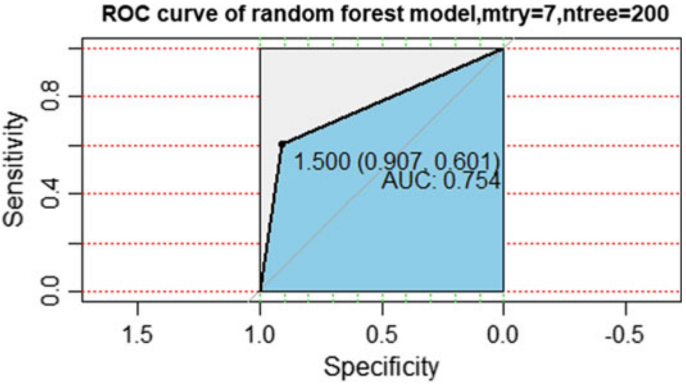


Fig. 3 ROC curve of delayed and default random forest model

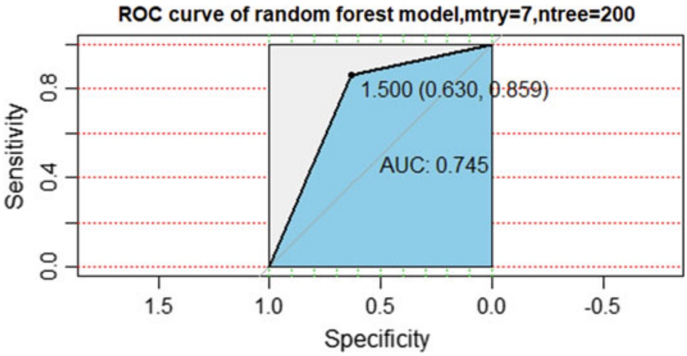


Fig. 4 ROC curve of normal and default stochastic forest model

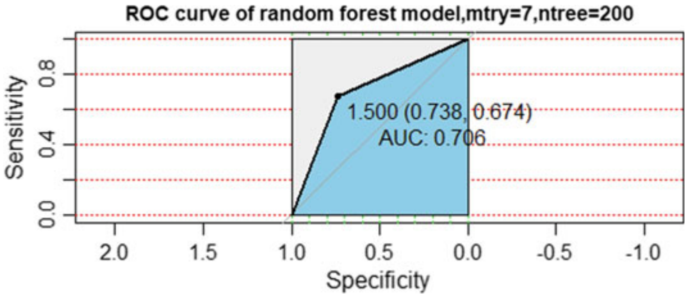


Fig. 5 ROC curve of normal and delayed random forest model

Table 2 Confusion matrix of three experiments

	Prediction			Prediction			Prediction	
Reference	0	1	Reference	1	2	Reference	0	2
0	471	29	1	329	171	0	422	78
1	203	297	2	73	427	2	148	352

Table 3 Model attribute importance sort

	Delay and charged-off	Fully paid and charged-off	Fully paid and delay
correct prediction	773	758	778
Error prediction	227	242	222
The percent of correct prediction (%)	77.3	75.8	77.8

It can be seen from Table 3 that the prediction accuracy of the three models is reduced, because the size of the training set data is large, while the size of the prediction sample is small. As the number of samples increases, the prediction accuracy will increase. Another reason is that to avoid the impact of unbalanced data, the number of samples of normal repayment is the same as that of delay and default when modeling, so there is an error between the number of normal repayment and the number of delay or default predicted, which depends on more experiments to verify. The model still has a good prediction effect, which is not different from the theoretical model, indicating that the model has a strong generalization ability.

Table 4 shows the importance of the attributes of the random forest model. From the results of attribute importance ordering, different models have different attribute importance ordering, which reflects the different contribution rates of attributes to the model under different target variables.

It can be seen from the ranking of attributes in different prediction models that many attributes play an important role in different situations, such as out_prncp, emp_length, total_rec_prncp, total_rec_int, total_pymnt, int_rate, installation, grade, funded_amnt, revol_bal, bc_util, etc. What is different from Jina and Zhua [1] research is the repayment data, including ut_prncp, total_rec_prncp, total_rec_int, and total_pymnt are very important in all sorts, while grade and funded_amnt are less important than the repaid data, which can be interpreted as expected to become an important factor in repayment. The performance of emp_length in different classifications is quite different, which requires further study.

4 Conclusion

Due to the convenience of online credit loans, the extensive promotion of Internet companies and commercial banks has led to the rapid development of online loan products, and the number of online loan users has increased significantly. However, the credit risk of borrowers is still the main factor hindering the development of the personal online lending market. Therefore, it is necessary to promote the sustainable development of the online lending industry by mining the data of the online lending platform to study the influencing factors of the borrower's delinquency, and suggesting how to effectively recover the principal after the online platform borrows. This article selects the 2019 annual data of the Lending Club website, constructs the random forest models after data preprocessing, and analyzes the borrower's borrowing status, and draws the following conclusions:

(1) By constructing a random forest model for different independent variables and dependent variables, this paper finds the important attributes that affect whether to delay and default in normal customers, and the important attributes that affect whether to default in delayed customers.

Table 4 Importance ranking of attributes in different prediction models

Attribute	Attribute interpretation	Delayed and default importance	Normal and default importance	Normal and delayed importance
out_prncp	Remaining outstanding principal for total amount funded	1	1	2
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	2	N	1
total_rec_prncp	Principal received to date	3	2	3
total_rec_int	Interest received to date	4	5	8
total_pymnt	Payments received to date for total amount funded	5	3	6
int_rate	Interest Rate on the loan	6	7	4
installment	The monthly payment owed by the borrower if the loan originates.	7	4	5
grade	LC assigned loan grade	8	11	7
funded_amnt	The total amount committed to that loan now.	9	6	9
revol_bal	Total credit revolving balance	10	22	13
bc_util	Ratio of total current balance to high credit/credit limit for all bank card accounts.	12	12	10
total_bal_ex_mort	Total credit balance excluding Mortgage	14	8	11
total_bc_limit	Total bankcard high credit/credit limit	23	9	12
bc_open_to_buy	Total open to buy on revolving bankcards.	26	10	18

(2) By constructing a random forest model for the loan status of the lender, it can be found that the model can effectively predict the repayment situation of the lender after the loan, and thus provide a basis for the lender to take measures.

(3) In the ranking of the importance of the attributes of the random forest model, we found that for different loan statuses, the factors that need to be considered are different. How to find the most important factors among these influencing factors and reduce the supervision cost of online loan platforms has become the main direction of future research.

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant No. 71971025 and No. 71271027.

References

1. Y. Jin, Y. Zhu, A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending, in *The Proceeding of The Fifth International Conference on Communication Systems and Network Technologies*, p. 609–613 (2015)
2. J.J. Xu, Y. Lu, M. Chau, P2P lending fraud detection: a big data approach. Intell. Secur. Inf. **9074**, 71–81 (2015)
3. M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests. Expert Syst. Appl. **42**(10), 4621–4631 (2015)
4. L. Breiman, Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. M. Kumar, M. Thenmozhi, Forecasting stock index movement: a comparison of support vector machines and random forest, in *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, Working Paper Series (2006)
6. A. Montillo, H. Ling, Age regression from faces using random forest, in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Cairo, 7–10 Nov 2009, p. 2465–2468 (2009)
7. V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. **67**, 93–104 (2012)
8. K. Theofilatos, S. Likothanassis, A. Karathanasopoulos, Modeling and trading the EUR/USD exchange rate using machine learning techniques. ETASR—Eng. Technol. Appl. Sci. Res. **2**(5), 269–272 (2012)
9. Q. Qin, Q.G. Wang, J. Li, S.S. Ge, Linear and nonlinear trading models with gradient boosted random forest and application to the Singapore Stock Market. J. Intell. Learn. Syst. Appl. **5**, 1–10 (2013)
10. R. Emekter, Y. Tub, B. Jirasakuldech M. Lu, Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. Appl. Econ. **47**(1) 54–70 (2015). http://dx.doi.org/10.1080/00036846.2014.962222
11. A. Byanjankar, M. Heikkila J. Mezei, Predicting credit risk in peer-to-peer lending: a neural network approach, in *IEEE*, p. 719–725 (2015)

Stackelberg Game Model of Power Plants and Large Users Considering Carbon Trading

Xinyu Niu and Shifeng Liu

Abstract Bilateral transactions are widely used in the power market in the context of power reform and low-carbon power, and more and more users can directly sign bilateral contracts with power plants to purchase power. This paper studies the issue of bilateral transactions between multiple power generators and large users based on the carbon trading background. The power generators first gave a quotation, and the large users then decided the stackelberg game problem of the contract power. The results show that the existence of a Nash equilibrium solution for multiple units quoting at the same time, and an algorithm for determining the contract power of large users and the initial price of units is given. This study of strategic behavior can provide theoretical support for decision-making of power plants and large users, and provide theoretical support for the bilateral negotiation between the unit and large users.

Keywords Spot electricity market · Bilateral transaction · Stackelberg game · Carbon trading

1 Introduction

In 2015, a new round of power reform which starting from the approval of transmission and distribution prices, promoting direct power trading, opening the power supply market at both ends gradually and forming electricity trading market was clarified in the *Opinions of the Central Committee of the Communist Party of China and the State Council on Further Deepening the Reform of the Electric Power System.* With the release of competitive electricity prices or the release of planned electricity generation and electricity consumption to a certain percentage, or the deviation of

X. Niu · S. Liu (✉)
School of Communication and Information, Beijing Jiaotong University, Beijing, China
e-mail: shfliu@bjtu.edu.cn

X. Niu
e-mail: 14113125@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_58

contract execution power cannot be resolved in accordance with the methods stipulated in this rule, localities should start the construction of electricity spot markets and establish mid-to-long-term trading and Market-based electricity and electricity balance mechanism combining spot transactions. As an important part of China's electricity marketization reform, the implementation of direct electricity purchase transactions by large customers marks that China's electricity market has begun to shift from a single power generation market to a bilateral market with access to the load side. According to the nature of medium- and long-term contracts, the essential purpose of bilateral transactions is to lock in trading prices and trading power in advance, and try to avoid the risks faced by market entities in the spot market.

With the intensification of global warming, the power industry, as the main industry for fossil fuel consumption, accounts for nearly 40% of the total carbon emissions from energy consumption. Carbon trading, as a market-oriented means to promote carbon emission reduction, has been widely used in various countries around the world. In 2013, China launched carbon trading pilots in 7 provinces, including Beijing and Shanghai. In December 2017, the *National Development and Reform Commission issued the National Carbon emission trading market construction plan (power generation industry)* signifies that China's carbon emission trading system is the first to start with the power generation industry as a breakthrough. The construction and improvement of the carbon trading market will be an important part of China's low carbon development in recent years.

In the case of considering carbon trading, in the bilateral transactions in which the power producers and large power users directly negotiate face-to-face, both parties in the transaction hope to optimize their utility functions through the formulation of their own strategies. However, since the power producer and the large user are two subjects with conflicting interests, the formulation of one party's strategy will directly affect the utility of the other party, so there is a problem of interest game between the two parties in the transaction. Based on the bilateral trading model of mid-term and long-term electricity trading, this paper establishes a game model for generating units and power users that considers carbon trading, and analyzes how power plants and large users make decisions and how to sign bilateral contracts.

2 Literature Review

In the power market environment, strategic bidding is one of the important decision-making behaviors of market participants. Each market participant achieves the purpose of maximizing its own profits by quoting to the ISO (independent system operator) strategy of the power dispatching center. The research on market participants' strategic bidding behavior can provide market participants with decision-making reference. At the same time, the analysis of the characteristics of market equilibrium points based on strategic bidding is also of great significance for the exploration and design of market mechanisms. The current research on market participants' strategic bidding is mainly based on theories such as game theory [1], multi-agent simulation platform [2], and experimental economics [3]. The literature [1] based on the game theory establishes a master-slave game model of bilateral contract transactions between multiple power producers and multiple large users, which proves the existence of Nash equilibrium solutions for multiple power generators quoting at the same time, and gives the main solving steps from the game model. Reference [2] solves the bidding strategy of power producers and the purchasing strategy of power consumers by constructing a master-slave game model that takes into account bilateral contract transactions between multiple power producers and multiple large users. Bilateral transactions refer to power generation enterprises and large users who meet the entry conditions to determine the transaction power and transaction price through independent negotiation, and after the security institutions have verified the contract, they will sign direct transaction power purchase and sales contracts. The advantage of bilateral transactions is to give full play to the subjective initiative of market players, and strive to maximize their own interests through negotiation, while the disadvantage is that there is more serious information symmetry between the power generation companies and users, the negotiation cycle is long, and the transaction cost is high. So, bilateral transactions are suitable for medium and long-term direct transactions and market entities with strong bargaining power. Besides, literature [4] studied how the two parties to purchase and sell electricity formulate the optimal bilateral contract under the background of a cooperative game between the electricity retailer and the power generator. What's more, literatures [5] and [6] studied the electricity retailer's electricity purchase strategy in the spot market and bilateral contract market and its strategy to sell electricity to residential customers, established a two-tier game model of the retailer's electricity purchase and sale profit, and gave out of their respective Nash equilibrium solutions. However, the above-mentioned literature basically studies bilateral transactions between power plants and large users, without considering the impact of carbon trading background on the decisions of power plants and large users.

In terms of considering carbon trading, in literature [7], the dispatching center revised its quotation according to the actual CO_2 emission intensity of the generating units, so that the units adopting advanced emission reduction technology were given priority, promoting system emission reduction; Reference [8] proposes a low-carbon dispatch optimization model for power generators that considers carbon trading risks; Reference [9] studies the demand side reserve bidding and scheduling issues under the carbon emissions trading system; Reference [10] Based on the classic Cournot game competition model, the strategic bidding behavior of generators considering carbon cost was studied. In [11], a two-layer optimization model was established to explore the framework of a composite carbon emission reduction policy considering carbon trading and carbon tax. Optimize the decision-making process of carbon emission reduction between government and enterprises. Literature [12–14] based on the bilateral bidding model of the electricity market, introduced a carbon trading mechanism, studied the strategic competitive price game of power producers and power users considering carbon trading, and established a two-level optimization model of strategic bidding for power generators and power users [15]. The diagonal equilibrium iterative algorithm is used to solve the game equilibrium point [16, 17].

Based on the simultaneous decision of the power producer and the large user, a two-level optimization model of the strategic bidding of the power producer and the power user is established.

Based on the above researches, this paper studies the issue of bilateral transactions between multiple power generators and large users based on the carbon trading background.

The power generators and large users do not make decisions at the same time, but the stackelberg game problem where the power generators first give a quotation, and then the large users decide the contract power.

3 Stackelberg Game Model of Power Plant Units and Large Users Considering Carbon Emissions

This paper studies bilateral trading scenarios between multiple units and multiple large users and assumes that a bilateral contract transaction lasts T cycles, and there are I units and J large users participating in the transaction. In the initial stage of contract signing, the unit's $i(i = 1, 2, \ldots, I)$ contract offer for large users $j(j = 1, 2, \ldots, J)$ is $\left(a_{ij}, b_i\right)$, including the initial price a_{ij} of the contract and the growth factor $b_i > 0$ of the contract price for the contracted electricity. Therefore, in the cycle $t(t = 1, 2, \ldots, T)$, if the power contract signed between the large user j and the unit i is q_{ij}^t, the contract price is $p_{ij}^t = a_{ij} + b_i q_{ij}^t$. The electricity demand of large users j in each cycle is $\left(\mathrm{D}_j^1, \mathrm{D}_j^2, \ldots, \mathrm{D}_j^\mathrm{T}\right)$, and the unit electricity benefit in each cycle is $\left(u_j^1, u_j^2, \ldots, u_j^\mathrm{T}\right)$. Large users $j(j = 1, 2, \ldots, J)$ predict that the spot price of electricity purchased in the electricity spot market will be $\left(p_s^1, p_s^2, \ldots, p_s^T\right)$. Large customers $j(j = 1, 2, \ldots, J)$ make the most profits by making decisions on the contracted power $\left(q_{ij}^1, q_{ij}^2, \ldots, q_{ij}^T\right)$ with the unit $i(i = 1, 2, \ldots, I)$ and the purchase of electricity $\left(q_{sj}^1, q_{sj}^2, \ldots, q_{sj}^T\right)$ in the spot market. These units are in a competitive relationship. The unit's $i(i = 1, 2, \ldots, I)$ unit power generation cost in each cycle is $\left(c_i^1, c_i^2, \ldots, c_i^T\right)$. The unit's i cyclical carbon emission right during cycle t is proportional to its contribution to the system's carbon emissions without carbon emission restrictions, where n is the unit electricity emission allocation, e_i is the unit's i carbon emission intensity and q_i^t is the unit's i contribution of J large users countermeasures without carbon emissions restrictions, so the unit's i carbon emission right in each cycle is $\left(ne_i q_i^1, ne_i q_i^2, \ldots, ne_i q_i^T\right)$. When the actual carbon emission of the unit is less than the allocated carbon emission quota, the remaining credit can be sold in the carbon trading market for profit; and when the actual carbon emission of the power generator is greater than the allocated carbon emission quota, there is a need to go the carbon trading market purchases the excess, otherwise the power producer will face a high fine. This study assumes that the sum of the actual carbon emissions of each unit does not exceed the system carbon emission limit, that is, high fines are not considered. The market carbon price for each cycle is a uniform price ω. As the unit's i contract power price has a fixed growth factor b_i with respect to contract power, the unit i will choose its own contract's initial price quotation strategy a_{ij} by predicting the power purchase strategy of large customers to maximize its own profits.

The parameters are set as follows (Table 1).

According to the above analysis, it can be known that the unit first gives a contract quotation, and then the large user decides the contract power. Therefore, the unit and the large user are the stackelberg game in which the unit decides first and the large user decides later.

First, analyze the decision-making behavior of large users $j(j = 1, 2, \ldots, J)$: After the large users j obtain the unit's $i(i = 1, 2, \ldots, I)$ contract quotation $\left(a_{ij}, b_i\right)$, the large users j according to the power demand D_j^t during the cycle t, the maximum power $\bar{q}_{ij}^t$ contracted by the large users j from the unit during the cycle t, the unit electricity benefit u_j^t during the cycle t, and the forecast of the spot price of electricity p_s^t in the spot market of electricity, optimize the total profit of large customers by deciding the period of contracted electricity q_{ij}^t signed with the unit i during t and the

Table 1 The parameters

Symbol	Meaning
a_{ij}	The initial electricity price of the unit's i contract for large users j
b_i	Coefficient of the unit's i contracted electricity price for large users j with respect to the contracted electricity growth factor
q_{ij}^t	Periodic contracted electricity signed by large users j and units i during t
$\bar{q}_{ij}^t$	During the period t, the maximum amount of contracted electricity obtained by the large users j from the unit i
p_{ij}^t	Contracted electricity price for large users j and units i during t,$p_{ij}^t = a_{ij} + b_i q_{ij}^t$
D_j^t	Large users' j power demand in the cycle t
u_j^t	Benefits of large users j in unit power consumption in a cycle t, $u_j^t > p_{ij}^t$ and $u_j^t > p_s^t$
p_s^t	Large users predict spot electricity prices purchased in the electricity spot market
q_{sj}^t	Large users' j electricity purchases in the spot market
c_i^t	Unit generating cost of unit i per cycle t
n	Unit electricity emission allocation
e_i	Carbon emission intensity of the unit i
q_i^t	Unit's i contribution to J large users without carbon emission restrictions
$ne_i q_i^t$	Carbon emissions right of the unit i in cycle t
ω	Market unit carbon price
Π_j	Large user's j total profit in T cycles
Π_j^t	Large user j profit in the cycle t
Π_{i}	Unit's i total profit in T cycles

amount of electricity q_{sj}^{t} purchased in the spot market. Thus, the contracted electricity price signed between the large user j and the unit i during t is $p_{ij}^{t} = a_{ij} + b_i q_{ij}^{t}$.

The total profit for large users j is:

$$\begin{aligned}
&\prod_j \left(q_{1j}^{1}, q_{2j}^{1}, \ldots, q_{Ij}^{1}, q_{sj}^{1}, \ldots, q_{1j}^{T}, q_{2j}^{T}, \ldots, q_{Ij}^{T}, q_{sj}^{T}\right) \\
&\quad= \sum_{t=1}^{T}\left(u_j^t\left(\sum_{i=1}^{I} q_{ij}^{t} + q_{sj}^{t}\right) - \sum_{i=1}^{I} p_{ij}^{t} q_{ij}^{t} - p_s^t q_{sj}^{t}\right) \\
&\quad= \sum_{t=1}^{T}\left(u_j^t\left(\sum_{i=1}^{I} q_{ij}^{t} + q_{sj}^{t}\right) - \sum_{i=1}^{I} \left(a_{ij} + b_i q_{ij}^{t}\right) q_{ij}^{t} - p_s^t q_{sj}^{t}\right) \\
&\quad= \sum_{t=1}^{T}\left(\sum_{i=1}^{I} \left(\left(u_j^t - a_{ij}\right) q_{ij}^{t} - b_i q_{ij}^{t2}\right) + \left(u_j^t - p_s^t\right) q_{sj}^{t}\right)
\end{aligned} \tag{1}$$

Optimize the total profit of large users j as:

$$\begin{aligned}
&\max_{q_{1j}^{1},\ldots,q_{Ij}^{1},q_{sj}^{1},\ldots,q_{1j}^{T},\ldots,q_{Ij}^{T},q_{sj}^{T}} \Pi_j\left(q_{1j}^{1}, q_{2j}^{1}, \ldots, q_{Ij}^{1}, q_{sj}^{1}, \ldots, q_{1j}^{T}, q_{2j}^{T}, \ldots, q_{Ij}^{T}, q_{sj}^{T}\right) \\
&s.t.\ \ 0 \le q_{ij}^{t} \le \bar{q}_{ij}^{t} \quad i = 1, 2, \ldots, I,\ \ t = 1, 2, \ldots, T \\
&\sum_{i=1}^{I} q_{ij}^{t} + q_{sj}^{t} = \mathrm{D}_j^t \quad t = 1, 2, \ldots, T \\
&q_{sj}^{t} \ge 0 \quad t = 1, 2, \ldots, T
\end{aligned} \tag{2}$$

Among them, the objective function $u_j^t\left(\sum_{i=1}^{I} q_{ij}^{t} + q_{sj}^{t}\right)$ is the power consumption benefit of large users j in the cycle t, $p_{ij}^{t} q_{ij}^{t}$ is the power purchase cost of large users j buying power from the unit i in the cycle t, and $p_s^t q_{sj}^{t}$ is the predicted power purchase cost of large users j in the cycle t. The first constraint in the optimization model (2) indicates that during the period t, the contract power q_{ij}^{t} signed between the large user j and the unit i is not less than zero, and does not exceed the upper limit of the contract power $\bar{q}_{ij}^{t}$ obtained by the large user j from the unit i. The second constraint indicates that the sum of the periodic contracted electricity signed by each large user j with each unit i and the spot purchased electricity in the spot market in the cycle t is equal to the large user's j electricity demand during the cycle t, which is the supply-demand balance in large users. The third constraint indicates that large users j purchase no less than zero electricity in the spot market during the cycle t.

This article assumes that the electricity spot market electricity price has nothing to do with the power purchase strategy of large users, and assumes that the unit has an independent quotation curve for each major user, that is, the final contract price of the unit to each major user is only affected by the major user's power purchase strategy. Therefore, there is no game between large users on the contracted power in each unit, and the power purchase strategies of large users are independent of each other.

Then analyze the unit's i decision-making behavior: In the case that other units' bidding strategies remain unchanged, the unit i decides the bidding strategy by predicting the contracted electricity q_{ij}^t amount signed by large users j with the unit i during the period t, the unit i is based on the unit generation cost c_i^t in the cycle t, the unit's carbon emission rights $ne_i q_i^t$ in the cycle t, the market unit carbon price ω, and the growth coefficient b_i of the unit's i contract electricity price. The unit i decides the initial electricity price of the unit's contract a_{ij} with large users j through prediction, and optimizes the unit's total profit. Thus, the quotation strategy of the large user j and the contracted electricity price $p_{ij}^t = a_{ij} + b_i q_{ij}^t$ signed by the large user j and the unit i in the cycle t was determined.

The total profit of the unit i is:

$$\begin{aligned}
&\prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ}) \\
&\quad = \sum_{t=1}^{T} \left(\sum_{j=1}^{J} p_{ij}^t q_{ij}^t - c_i^t \sum_{j=1}^{J} q_{ij}^t - \omega(1-n) e_i q_i^t \right) \\
&\quad = \sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left(p_{ij}^t - c_i^t\right) q_{ij}^t - \omega(1-n) e_i q_i^t \right) \qquad (3)
\end{aligned}$$

The total profit of the optimized unit i is:

$$\begin{aligned}
&\max_{a_{i1}, a_{i2}, \ldots, a_{iJ}} \prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ}) \\
&s.t. \quad \sum_{j=1}^{J} q_{ij}^t = q_i^t \quad t = 1, 2, \ldots, T \\
&a_{ij} \geq 0 \quad j = 1, 2, \ldots, J \qquad (4)
\end{aligned}$$

The $p_{ij}^t q_{ij}^t$ in the objective function is the electricity sales revenue obtained from large users j in the cycle t, $c_i^t \sum_{j=1}^{J} q_{ij}^t$ is the generation cost of the cycle t, and $\omega(1-n)e_i q_i^t$ is the carbon emission cost of the cycle t. When $n < 1$, the actual carbon emissions of the unit i is greater than the carbon emission quota, and the excess amount need to purchase in the carbon trading market. When $n > 1$, the actual carbon emissions of the unit i at that time were less than the allocated carbon emission quota, the remaining credits could be sold in the carbon trading market for profit. When $n = 1$ the actual carbon emissions of the unit i at that time were equal to the allocated carbon emission quota, and the morning carbon emissions report is zero. The first constraint in the optimization model (4) indicates that the total contracted power of the unit i signed with each large user j is equal to the unit's i

contribution to large users j without carbon emission restrictions, that is, the unit's i supply and demand balance. The second constraint indicates that the initial price of the unit's i contract in the cycle t is not less than zero.

This article studies maximizing the unit's revenue in the bilateral contract market, and therefore ignores the electricity sales profit of the power generator in the spot market when modeling. When it is necessary to sell electricity in the spot market, the unit can optimize the decision in the spot market according to the bilateral contracts already signed. This article assumes that the game between power producers is a non-cooperative static game with complete information, and the game between power producers and large users is a non-cooperative dynamic game with complete information. Therefore, all game participants know not only their own strategy space and profit function, but also the strategy space and profit function of other participants.

4 Model Analysis and Solving

From the models (1) to (2), we can know that the power purchase optimization problem for large users can be regarded as a set of optimization problems individually in each cycle. The optimization problem of each cycle is as follows

The profits of large users j in the cycle t are:

$$\begin{aligned}&\prod_{j}^{t}\left(q_{1j}^{t}, q_{2j}^{t}, \ldots, q_{Ij}^{t}, q_{sj}^{t}\right)\\ &\quad=\sum_{i=1}^{I}\left(\left(u_{j}^{t}-a_{ij}\right)q_{ij}^{t}-b_{i}q_{ij}^{t2}\right)+\left(u_{j}^{t}-p_{s}^{t}\right)q_{sj}^{t}\end{aligned} \tag{5}$$

Optimize the profits of large users j in the cycle t:

$$\begin{aligned}&\max_{q_{1j}^{t},q_{2j}^{t},\ldots,q_{Ij}^{t},q_{sj}^{t}}\prod_{j}^{t}\left(q_{1j}^{t}, q_{2j}^{t}, \ldots, q_{Ij}^{t}, q_{sj}^{t}\right)\\ &s.t.\quad 0\le q_{ij}^{t}\le \bar{q}_{ij}^{t}\quad i=1,2,\ldots,I,\\ &\sum_{i=1}^{I}q_{ij}^{t}+q_{sj}^{t}=\mathrm{D}_{j}^{t}\\ &q_{sj}^{t}\ge 0\end{aligned} \tag{6}$$

From $b_i > 0$ and $u_j^t > p_{ij}^t$, it is clear that models (5) to (6) are a strictly concave quadratic programming problem with a non-empty feasible solution set. According to LUO (1996), Lemma 1.

Theorem 1 *The optimal solutions* $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ *and* $q_{sj}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ *of models* (5) *to* (6) *are unique for any of* a_{ij}. *And* $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ *is a piecewise smooth linear function on* $\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$.

The KKT conditions of models (5) to (6) are

$$\begin{cases} 2b_i q_{ij}^t - \left(u_j^t - a_{ij}\right) + \alpha_i + \beta = 0 & i = 1, 2, \ldots, I \\ p_s^t - u_j^t + \beta = 0 & \\ \alpha_i\left(q_{ij}^t - \bar{q}_{ij}^t\right) = 0 & i = 1, 2, \ldots, I \\ \sum_{i=1}^{I} q_{ij}^t + q_{sj}^t = \mathrm{D}_j^t & \\ q_{ij}^t \geq 0, q_{sj}^t \geq 0 & i = 1, 2, \ldots, I \end{cases} \tag{7}$$

Therefore, the system of Eq. (7) is solved to obtain the only optimal solution $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ and $q_{sj}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$.

The total profit of the unit i is:

$$\prod_i (a_{i1}, a_{i2}, \ldots a_{iJ}) = \sum_{t=1}^{T}\left(\sum_{j=1}^{J}\left(p_{ij}^t - c_i^t - \omega(1-n)e_i\right)q_{ij}^t\right) \tag{8}$$

The total profit of the optimized unit i is:

$$\max_{a_{i1}, a_{i2}, \ldots, a_{iJ}} \prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ})$$
$$s.t. \quad a_{ij} \geq 0 \quad j = 1, 2, \ldots, J \tag{9}$$

Substitute the optimal contract power of large users j for contract quotes and spot prices $q_{ij}^{t*}\left(a_{1j}, \cdots, a_{Ij}, p_s^t\right)$ to models (8) to (9), they can be simplified as follows:

The total profit of the unit i is:

$$\prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ}) = \sum_{t=1}^{T}\left(\sum_{j=1}^{J}\left(p_{ij}^t - c_i^t - \omega(1-n)e_i\right)q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)\right) \tag{10}$$

The total profit of the optimized unit i is:

$$\max_{a_{i1}, a_{i2}, \ldots, a_{iJ}} \prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ})$$

$$s.t. \quad a_{ij} \geq 0 \quad j = 1, 2, \ldots, J \tag{11}$$

From Theorem 1, we know that $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ is a piecewise smooth linear function, so $\prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ})$ is also a piecewise smooth function about $\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$.

Similar to MZERSON (1991) and Wu et al. [1], the existence of the equilibrium solution is known as Theorem 2.

Theorem 2 *When the contract power of large users j is* $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$*, break-point* ω_{ij} *satisfying* $\prod_i \left(\omega_{ij}, a_{i1}, a_{i2}, \ldots, a_{iJ}\right) > 0$ *for any* $\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$*, if the following formula is satisfied, at least one of the models* (10) *to* (11) *is a Nash equilibrium solution.*

$$\lim_{x^+ \to \omega_{ij}} q_{ij}^{t*}\left(x, a_{(-i)j}, p_s^t\right) \geq \lim_{x^- \to \omega_{ij}} q_{ij}^{t*}\left(x, a_{(-i)j}, p_s^t\right)$$

Among them $a_{(-i)j=}\left(a_{1j}, \ldots, a_{i-1j}, a_{i+1j}, \ldots, a_{Ij}, p_s^t\right)$.

Since KKT condition (7) satisfies Theorem 2, there is at least one model (10) to (11) that is a Nash equilibrium solution.

5 Algorithm Design

According to the analysis in Sect. 3, the following algorithm can be designed to obtain the unit's bidding strategy and the large user's order power strategy.

Step1. Initialize the parameters.

Step2. For the large user $j (j = 1, 2, \ldots, J)$, solve the KKT condition (7) in turn to get the large user's j contract power as $q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$.

Step3. Randomly select the initial power price a_{ij} of the contract of the generating unit $i (i = 1, 2, \ldots, I)$ within the feasible region.

Step4. Calculate the profit $\prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ})$ of the unit i at this time, and let the maximum profit be $\prod_i^* (a_{i1}, a_{i2}, \ldots, a_{iJ}) = \prod_i (a_{i1}, a_{i2}, \ldots, a_{iJ})$.

Step5. For the unit i, traverse the feasible region to find the solution $\left(a_{i1}^*, a_{i2}^*, \ldots, a_{iJ}^*\right)$ of the model (10) to (11) and the corresponding $\prod_i \left(a_{i1}^*, a_{i2}^*, \ldots, a_{iJ}^*\right)$, if $\prod_i^* \left(a_{i1}^*, a_{i2}^*, \ldots, a_{iJ}^*\right) > \prod_i^* (a_{i1}, a_{i2}, \ldots, a_{iJ})$ let $a_{ij} = a_{ij}^*$ and $\prod_i^* (a_{i1}, a_{i2}, \ldots, a_{iJ}) = \prod_i \left(a_{i1}^*, a_{i2}^*, \ldots, a_{iJ}^*\right)$.

Step6. Repeat step 5 until all the units do not modify the initial electricity price of their contracts and reach the Nash equilibrium. At this time, the initial electricity price a_{ij} of each unit is the initial electricity price of each unit under the Nash equilibrium solution.

Step7. For large users $j (j = 1, 2, \ldots, J)$, substitute the initial electricity price a_{ij} into the contracted power of the large user j, and let $q_{ij}^{t*} = q_{ij}^{t*}\left(a_{1j}, \ldots, a_{Ij}, p_s^t\right)$ to obtain the contracted power of the large user.

6 Conclusion

Bilateral transactions are widely used in the power market, and more and more users can directly purchase power through signing bilateral contracts with power plants. In this paper, a stackelberg game model considering carbon emissions was established in the context of bilateral negotiations between multiple units and multiple large users, and the existence of a Nash equilibrium solution for multiple units quoting at the same time was proved. The algorithm for solving the contract power and the initial electricity price of the unit provides theoretical support for the bilateral negotiation between the unit and the large user. On the one hand, in the game model of this paper, the generator's decision variable is its initial contract offer to different large customers, excluding the slope of the offer curve. The next work will further optimize the game model, so that the unit's decision variables include both the initial contract quote and the slope of the quote curve, and provide a method for solving the game model. On the other hand, this article assumes that the unit production cost of the unit is fixed, but the actual unit production cost is not fixed and a linear function is often used, that is, the production cost is a quadratic function. The next work will further optimize the game model and analyze the impact of different unit cost parameters on decision-making.

References

1. C. Wu, B. Gao, T. Yi et al., Master-slave game based bilateral contract transaction model for generation companies and large consumers. Autom. Electr. Power Syst. **40**, 62 (2016)
2. Z. Jing, J. Zhu, Simulation experiment analysis on market rules for monthly centralized bidding. Autom. Electr. Power Syst. (24), 48–54 (2017)
3. Z. Hu, J. Xu, D. Gan, Design and analysis of auction experiment for electricity market. Autom. Electr. Power Syst. **028**(4), 10–16 (2004)
4. S.I. Palamarchuk, Bilateral contracts for electricity delivery: scheduling and arrangement, in *Proceedings of 2007 IEEE Lausanne Power Tech*. Lausanne, Switzerland, 1–5 July 2007, p. 843–848
5. W. Wei, F. Liu, S. Mei, Energy pricing and dispatch for smart grid retailers under demand response and market price uncertainty. IEEE Trans on Smart Grid. **6**(3), 1364–1374 (2015)
6. F.S. Oliveira, C. Ruiz, A.J. Conejo, Contract design and supply chain coordination in the electricity industry. Eur. J. Oper. Res. **227**(3), 527–537 (2013)
7. M. Bosnjakovic, M. Stojkov, M. Jurjevic, Environmental impact of geothermal power plants. Tehnicki vjesnik-Technical Gazette **26**(5), 1515–1522 (2019)
8. E. Comăniță, P. Cozma, I. Simion, M. Roșca, M. Gavrilescu, Evaluation of eco-efficiency by multicriteria decision analysis. Case study of eco-innovated and eco-designed products from recyclable waste. Environ. Eng. Manag. J. **17**, 1791–1804 (2018)
9. S. Bharwana, S. Ali, M. Farid, M. Zubair, M. Rizwan, R. Ahmad, Occupational health and safety conditions in small medium sized enterprises of iron furniture manufacturing units. Environ. Eng. Manag. J. **18**, 545–553 (2019)
10. X. Li, X. Cai, C. Fu, An energy saving and emission reduction based bidding transaction mode under carbon trading mechanism. Autom. Electr. Power Syst. **35**(10), 48–52 (2011)
11. M. Sun, L. Liu J. Yu, et al., Low carbon power dispatch strategy for power suppliers considering the risks of carbon trading. Southern Power Syst. Technol. **11**(11), 46–52 (2017)

12. X. Liu, X. Ai, J. Yang, Design of demand side reserve bid-scheduling strategy considering future carbon emission trading. Autom. Electr. Power Syst. **35**(2), 38 (2011)
13. J. Li, Y. Chen, S. Liu et al., Electricity market equilibrium analysis considering carbon emission cost. Power Syst. Technol. **40**(5), 1558–1563 (2016)
14. Y. Hao, C. Tian, C. Wu, Modelling of carbon price in two real carbon trading markets. J. Clean. Prod. **244**, 118556 (2020)
15. H. Lu, X. Ma, K. Huang et al., Carbon trading volume and price forecasting in China using multiple machine learning models. J. Clean. Prod. **249**, 119386 (2020)
16. D. Stănciulescu, C. Zaharia, Process water treatment in a thermal power plant: characteristics and its sediment/sludge disposal. Environ. Eng. Manag. J. **19**, 255–267 (2020)
17. Y.J. Xiao, Y. Zheng, L.M. Zhang, Y.H. Kuo, A combined zone-LP and simulated annealing algorithm for unequal-area facility layout problem. Adv. Prod. Eng. Manag. **11**(4), 259–270 (2016)

Collaborative Governance of Internal Control of Scientific Research Funds in Colleges and Universities Under the Background of “Release Management Service”

Tian Qi, Xue-wei Li, and Jing Li

Abstract The research which starts from the national “Release Management Service” policy, and is guided by the national internal control documents, and also is based the COSO five-element classic framework, innovatively establishes the model of the internal control optimization of university scientific research funds and of the internal control dynamic circulation collaborative governance of university scientific research funds. On the basis of the cyclic collaborative governance model, the evaluation index system for the collaborative control of scientific research funds internal control in colleges or universities is established and will be applied to Z University. On the whole, the optimal path for collaborative management of scientific research funds internal control has been worked out to expand the theoretical model and provide the quantified practical exploration for the internal control under this policy.

Keywords Release management service · Research funds · Collaborative governance · Governance · Internal control · Colleges and universities

1 Introduction

According to the statistics compiled by the Science and Technology Department of the Ministry of Education from 2008 to 2018, the investment amount of science and technology funds in various colleges and universities was 177.279 billion RMB in 2018, which is 3.25 times higher than the 45.533 billion RMB in 2008. The rapid growth of scientific research funds in colleges and universities has increased the difficulty of internal control. According to Table 1, the state and the Ministry of Education have continuously revised and improved the control of scientific research funds

T. Qi · X. Li (✉) · J. Li
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: xueweili@bjtu.edu.cn

T. Qi
e-mail: 17113134@bjtu.edu.cn

J. Li
e-mail: jingli@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_59

Table 1 National level "release management service" research fund reform documents

No.	Publishing department	Symbol	Files
1	State Council	Guofa [2014] No. 11	Opinions on Improving and Strengthening Central Financial Research Projects and Fund Management
2	Ministry of Science and Technology Ministry of Finance	Caijiao [2015] No. 154	Notice on Issues Concerning Fund Management in the Transitional Period of Central Fiscal Science and Technology Plan Management Reform
3	CPC Central Committee and State Council	Development Office [2016] No. 50	Several Opinions on Further Improving the Fund Management Policy of Central Fiscal Science Research Projects
4	Ministry of Finance Ministry of Science and Technology	Finance Science Education [2016] No. 113	Measures for the Administration of Funds for National Key R&D Programs
5	Ministry of Finance Ministry of Science and Technology Ministry of Education Development and Reform Commission	Finance Science and Education [2017] No. 6	Further Improve the Implementation of Policies such as the Management of Funds for Central Government Research Projects
6	Ministry of Education	Religion and Politics [2017] No. 7	Several Opinions of the Five Departments including the Ministry OF Education on Deepening the Reform of Reduction, Decentralization, and Decentralization of Higher Education and Optimizing Service Reform
7	State Council Office	State Council Issued [2018] No. 127	Notice on Doing a Good Job in Implementing and Optimizing Documents to Give Scientific Institutions and Staff More Autonomy
8	Ministry of Science and Technology Ministry of Finance	National Science Office Funding [2018] No. 122	Notice on Carrying out Relevant Work of Resolving Scientific Research Funds
9	State Council	Guofa [2018] No. 25	Notice of the State Council on Optimizing the Management of Scientific Research Funds and Improving Scientific Research Performance

(continued)

Table 1 (continued)

No.	Publishing department	Symbol	Files
10	Ministry of Education	Church Party Letter [2019] No. 37	Notice of the Party Group of the Ministry of Education of the People's Republic of China on Doing a Good Job in Implementing Relevant Documents Giving More Autonomy to Scientific Research Management

in colleges and universities. The [Number 50] document issued in 2016 provides national policy guidance for the internal control of scientific research funds. It has become a new requirement for colleges and universities to combine "service" and "release management" closely to effectively enhance the innovation vitality of scientific researchers and to improve the efficiency of scientific and technological output. Therefore, it will be a great practical significance to explore the construction of internal control and collaborative governance of scientific research funds in colleges and universities for the implementation of "Release Management Service" policy.

2 Support of Internal Control Theory

As early as the 1930s, the prototype of internal control-internal checks and balances has attracted much attention from management scientists. With the development of social economy and the maturity of enterprise management, the theory of internal control has been derived. The release of the COSO report in 1992 was the peak of the rapid development of internal control theory, Its mainstream view is "the implementation of actions at all levels of the subject, which is intended to implement its main objectives and provide a reasonable assurance process" [1]. The research on the internal control of scientific research funds in colleges and universities at home and abroad mostly adopts the COSO five-element "control environment, control activities, risk assessment, information communication, supervision and evaluation" internal control framework as the basic theory of the research, see Fig. 1.

"Accounting Law" promulgated in China in 2000 was the first internal control-related regulations in China. In 2008, the Ministry of Finance promulgated the "Basic Standards for Internal Control of Enterprises". Scholars have explored the internal control plan for scientific research funding of universities and proposed corresponding solutions based on the "Guide to Internal Control of Economic Activities of Universities Directly under the Ministry of Education" and the five elements of the COSO framework. The research mainly focused on theoretical research and did not form a universality. The internal control model of scientific research funding of universities shows a large research space.

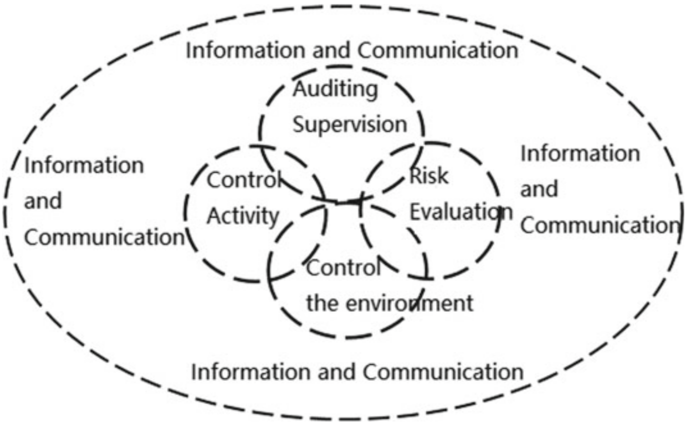

Information and Communication
Auditing Supervision
Information and Communication
Control Activity
Risk Evaluation
Information and Communication
Control the environment
Information and Communication

Fig. 1 A classic model of internal control of scientific research funding based on the COSO framework

3 Support of Collaborative Governance Theory

The core concept of collaborative governance is collaborative cooperation. At present, domestic academic circles mostly regard the cross combination of "synergetic" in natural science and "governance theory" in social science as the definition of collaborative governance. The combination of the two can produce "1 + 1 > 2" effect. In 2004, Harvard University Scholar Dohahue, one of the research authorities on collaborative governance, put forward to the concept of "collaborative governance" in the field of public management, which is a process of distributing or sharing discretionary power to subjects other than the government and realizing win-win cooperation in order to pursue the public goals designated by the government [2]. Collaborative governance has gradually become a new governance paradigm. The premise assumption is the "multi-center" theory, which is based on trust and adheres to the power dimension of up-down interaction and the authoritative source of resource integration. Its fundamental purpose is to realize the transformation from traditional governance paradigm to the collaborative one [3]. Synergetic and governance theories merge to form a cross-cutting theory category of multi-governance subject ordering-collaborative governance theory. Its essence lies in introducing the theory of systematic phase change in natural science into the theory of social governance, thus providing a new research perspective and method for social science research.

Synergy theory was put forward by German physicist h harken in 1971 [4]. He believes that coordination can realize the stable and orderly evolution of the whole process by controlling the order parameters in the whole process of activities. In the process of development and evolution of the system, cooperation among subsystems will produce a new orderly and stable structure. In 1995, the global governance Committee put forward the concept of governance theory: "the sum of the methods used by society, individuals and various public and private institutions to manage their same affairs is a continuous process in which conflicting or different stakeholders coordinate differences and join forces" [5]. It is very necessary to introduce the concept of cooperative governance on the basis of internal control of scientific research funds in colleges and universities to improve the use efficiency of funds. According to the characteristics of the research object of the cooperative governance theory, the accounting theory can be moved forward, integrated and coordinated. Thus clarifying the responsibility and authority of the cooperation between relevant scientific research departments and the efficient integration of supervision and control. Therefore, the introduction of this concept will certainly provide new theoretical basis for the expansion of the internal control theory and the solution of the serious process fragmentation and low quality benefit in this field.

4 Construction of Internal Control Model for Scientific Research Funding of Universities Based on "Decentralized Service" Under the COSO Framework

Reform of "delegation management service" basically lies in strengthening the use of funds, improving the internal control system, innovating supervision methods, and strengthening the building of integrity. Following the construction principles of "comprehensive, important, check and balance, and adaptation", the "decentralization" is based on the classic model of internal control in Chap. 2. The model of internal control of scientific research funding in the context of "management service" is shown in Fig. 2.

This model is constructed with different graphics to emphasize the important role played by the five elements of internal control. The bottom triangle represents the control environment and is the cornerstone and prerequisite for achieving internal control; the gears and arrows represent control activities and are the core and means of internal control; the internal circle represents risk evaluation is the basis and main line of internal control; the dotted outer circle and two z arrows represent the communication and the carrier of internal control, and can timely correct the deviation of internal control; the outermost pentagon represents the audit supervision, which is the focus and focus of internal control.

5 Construction of a Dynamic Cycle Model for Internal Control of University Scientific Research Funding Based on Collaborative Governance

Based on the concept of collaborative governance, guided by national internal control documents, and based on the COSO framework. This study attempts to build a SFIC

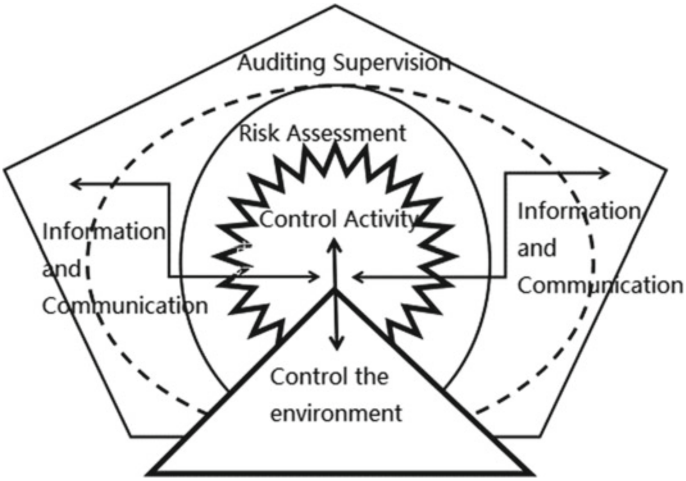


Fig. 2 Internal control model of scientific research funding of universities under the background of "delivery management suit"

model based on collaborative governance created by two scholars [6] under the influence of multiple external environments. A model for the dynamic and cyclical collaborative governance of internal control of university research funding is shown in Fig. 3.

This model is based on the domestic scholar Tian [7]'s universal model and the assumption that the two parts of the collaborative engine and collaborative behavior

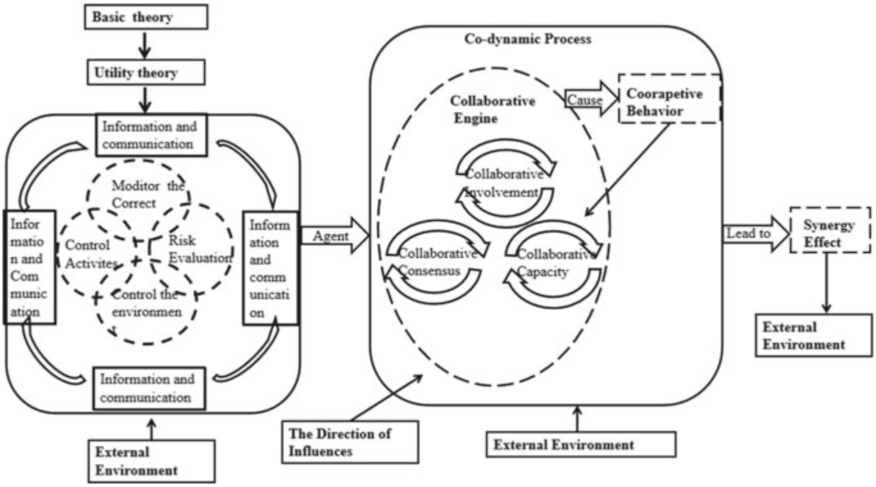


Fig. 3 A collaborative and dynamic governance model for internal control of scientific research funding in universities

are the most important parts of the collaborative process. It also draws on scholars Emerson and Nabatchi [8]. The newly established dynamic cycle model of internal control of university scientific research funding is replaced by ellipses to highlight the dynamics of collaborative governance, and cyclic arrows to replace gears to highlight the fluency and imperativeness of interaction between different elements. The new model is to replace cuboids with ellipses to highlight the dynamics of cooperative governance, to replace gears with circular arrows to highlight the fluidity of interaction between different elements, and to control the dynamic circulation model of scientific research funds in iterative colleges and universities.

In this research model, the collaborative engine is composed of collaborative consensus, collaborative participation, and collaborative capabilities to provide the driving force. It forms a dynamic cycle with collaborative behavior. It is the core part of this model. All collaborative departments form common goals based on the five elements of COSO. Through the sharing concept to guide collaborative behavior to produce synergistic consequences, it also stubbornly adapts to the external environment, thereby starting a specific full-process collaborative governance practice. When using this model, we need to understand several important notes: (1) Collaboration participation: Stakeholders involved in collaborative governance must consider both the overall internal control construction of the university and sufficient operational rights for specific operators. (2) Collaborative engine: in the process of collaborative participation and collaborative consensus promotes each other, the goal of internal control of scientific research funds in colleges and universities is the cooperative consensus of compliance and efficient use of scientific research funds assets. Only by reaching the positive interaction of cooperative consensus, cooperative participation and smooth change of cooperative ability can the engine exert itself to the cooperative behavior and lead to the cooperative dynamic cycle process. (3) Collaborative behavior: refers to the actions taken by the participating departments to implement the internal control objectives in the process of collaborative governance, but also allows for the immediate results of collaborative governance. (4) Collaborative consequences: through the impact, evaluation and accountability of relevant departments [9], the five goals of internal control collaborative governance are finally achieved: obtaining scientific research results, scientific and technological innovation, Obtain effective internal control efficiency, gain recognition from relevant external organizations or the public, and realize professional value. This study describes and studies the collaborative governance theory as a dynamic internal control process, and builds an interactive relationship between key factors to form an efficient benign internal control dynamic cycle process.

6 Construction and Application of Evaluation Index System for Collaborative Governance of Internal Control of Scientific Research Funds in Colleges and Universities

At present, there are qualitative and quantitative methods for evaluating the internal control of scientific research funds in colleges and universities. Qualitative evaluation methods include individual interviews, questionnaires, walk-through tests, comparative analysis, symposiums, etc. Individual interviews and questionnaires are suitable for understanding the status qua of internal control at the unit level in colleges and universities. Passing through test method, comparative analysis method and symposium method are suitable for internal control evaluation of economic and business activities in colleges and universities. However, due to the need for scientific evaluation of internal control, many scholars use AHP and fuzzy evaluation analysis to quantitatively evaluate the internal control results so as to facilitate the stakeholders and internal staff of scientific research funds in universities to use their evaluation results. Compared with the above-mentioned series of qualitative methods, analytic hierarchy process [10] has the advantage that abstract internal control evaluation objectives can be subdivided layer by layer into more intuitive and convenient evaluation indicators. The selected indicators for internal control elements tend to realize the combination of qualitative and quantitative, and build a multi-level and multi-index internal control evaluation system, which can more accurately realize the scientifically and hierarchy of internal control analysis for collaborative governance of scientific research funds in universities. Therefore, AHP is applicable to the evaluation of the effectiveness of internal control of the collaborative governance of scientific research funds in colleges and universities, which conforms to the systematic view and hierarchical relationship of AHP. It can not only set indexes for each component of internal control, but also assign different weight according to the characteristics of different units. It can realize comprehensive quantitative evaluation of multiple indexes and make up for the shortcomings of qualitative evaluation. Therefore, analytic hierarchy process is selected for this study.

In this paper, when constructing the evaluation index of internal control collaborative governance of scientific research funds in colleges and universities, based on the “Basic Evaluation Index and Scoring Scheme for Internal Control of Ministry of Education” and the “Basic Evaluation Standard System for Internal Control of Administrative Institutions” promulgated in 2016, and based on the COSO internal control framework, combined with the operation characteristics and actual situation of scientific research funds in colleges and universities, the relevant indexes of the “Basic Evaluation Index and Scoring Scheme for Internal Control of Ministry of Education” are divided in more detail. Referring to the research results of previous scholars [11], after consulting experts on the initially established indicators, the evaluation index system of internal control collaborative governance of university research funds with 5 primary indicators and 19 secondary indicators was finally determined, so as to more scientifically quantify the indicators to accurately evaluate its internal control effect, as shown in Fig. 4.

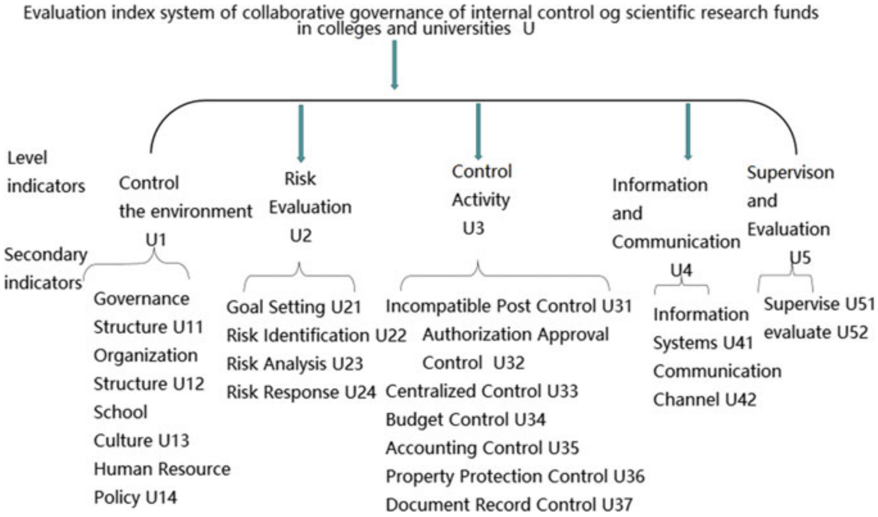

Evaluation index system of collaborative governance of internal control og scientific research funds in colleges and universities U
Level indicators
Control the environment U1
Risk Evaluation U2
Control Activity U3
Information and Communication U4
Supervison and Evaluation U5
Secondary indicators
Governance Structure U11
Organization Structure U12
School Culture U13
Human Resource Policy U14
Goal Setting U21
Risk Identification U22
Risk Analysis U23
Risk Response U24
Incompatible Post Control U31
Authorization Approval Control U32
Centralized Control U33
Budget Control U34
Accounting Control U35
Property Protection Control U36
Document Record Control U37
Information Systems U41
Communication Channel U42
Supervise U51
evaluate U52

Fig. 4 Evaluation index system for collaborative governance of internal control of scientific research funds in universities

In order to verify the scientific and effectiveness of the evaluation system, a subordinate Z university was selected to evaluate its internal control and collaborative governance of scientific research funding. Working with its scientific research project leaders and graduate students participating in scientific research projects, as well as scientific research, finance, assets, auditing and other departments, fifty questionnaires were sent out by the staff, all of them were recovered, and the pass rate was 100%. Evaluation index model U for the collaborative governance of scientific research internal control in universities was built-in the analytic hierarchy process MCE-AHP software, and the results are shown in Fig. 4. Specific evaluation indexes are U1, U2, …, and "Un" means that U = (U1, U2, …, Un). A judgment matrix for pairwise comparison is constructed, and the index comparison scale of the questionnaire is entered into each judgment matrix to establish each specific evaluation index "Un" for the comment. The membership matrix R of the set is as follows:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \quad \text{or } R = \left(r_{ij}\right), \quad i, j = 1, 2, \ldots, n$$

The questionnaire scoring results are substituted into the fuzzy evaluation method for data analysis. The membership degree vector b. The calculation formula is as follows:

Table 2 U's first-level index judgment matrix

U	U1	U2	U3	U4	U5
U1	1	1	1/5	3	3
U2	1	1	1/5	3	3
U3	5	5	1	7	7
U4	1/3	1/3	1/7	1	1
U5	1/3	1/3	1/7	1	1

$$\mathrm{B} = \mathrm{W} * \mathrm{R} = \begin{Bmatrix} r11 & r12 & \cdots & r1n \\ r21 & r22 & \cdots & r2n \\ \vdots & \vdots & \ddots & \vdots \\ rm1 & rm2 & \cdots & rmm \end{Bmatrix}$$

By analogy, the membership degree vector of the target layer u can be finally obtained, as showed in Table 2.

By calculation: the eigenvectors of the U matrix = (0.155, 0.155, 0.570, 0.06, 0.06), λmax = 5.0936. CI corresponding to the U matrix = 0.0234, CR = 0.0209 < 0.1, indicating that the judgment matrix passed the consistency test.

Finally, the evaluation index system weight and evaluation results are obtained, as showed in Table 3.

Evaluation scores for internal control and collaborative governance of scientific research funding of affiliated universities $\mathrm{D} = \{90, 70, 50, 30\}^{\mathrm{T}} * \{\mathrm{R}_1, \mathrm{R}_2, \mathrm{R}_3, \mathrm{R}_4\}$ = 80.73 points, which is "excellent" and highly consistent with the actual work. Through the construction of the evaluation index system of internal control and collaborative governance of scientific research funds in Z universities, Z universities have enhanced their understanding of the utilization rate and use of their own scientific research funds. It also shows that the internal control of scientific research funds in Z University is well coordinated. According to this evaluation result, the university can continue to improve the internal control of scientific research funds against the above internal control indicators, and can ensure the effective operation of its internal control to realize the sustainable development of scientific and technological innovation.

7 Analysis on the Achievements of Internal Control and Collaborative Governance of Scientific Research Funds in Z University

As the concrete implementer of the reform, under the background of "Release Management Service" policy and taking the internal control theory as the standard and holding the risk monitoring, the Z university has overcome the defects of weak

Table 3 Weights and evaluation results of the evaluation index system for internal control and collaborative governance of scientific research funding in universities

First-level indicators	Weights	Secondary indicators	Weights	Membership			
				Excellent	Good	Medium	Worse
Control environment	15.47%	Governance structure u11	3.87%	0.6	0.25	0.1	0.05
		Organization structure u12	3.87%	0.6	0.4		
		School culture u13	3.87%	0.3	0.5	0.2	
		Human Resources Policy u14	3.87%	0.5	0.5		
Risk assessment	15.47%	Goal setting u21	3.87%	0.4	0.5	0.1	
		Risk level u22	3.87%	0.5	0.5		
		Risk analysis u23	3.87%	0.5	0.5		
		Risk response u24	3.87%	0.6	0.4		
Control activity	57.04%	Incompatible post control u31	3.27%	0.5	0.4	0.1	
		Authorization approval control u32	3.27%	0.8	0.2		
		Centralized control u33	3.27%	0.3	0.5	0.2	
		Budget control u34	17.82%	0.7	0.3		
		Accounting control u35	17.82%	0.5	0.5		
		Property protection control u36	8.31%	0.6	0.4		
		File record control u37	3.27%	0.6	0.4		
Information and communication	6.01%	Information system u41	3.01%	0.5	0.4	0.1	
		Communication channel u42	3.01%	0.7	0.2	0.1	
Evaluation and supervision	6.01%	Supervision u51	3.01%	0.8	0.2		
		Evaluation u52	3.01%	0.7	0.3		
Evaluation summary			1	0.5716	0.3952	0.0313	0.0019
Comment score				90	70	50	30

(continued)

Table 3 (continued)

First-level indicators	Weights	Secondary indicators	Weights	Membership			
				Excellent	Good	Medium	Worse
Judging score			80.73	51.44	27.66	1.57	0.06

control environment, lax control activities, weak risk assessment, poor information communication, insufficient audit knowledge in the internal control of scientific research funds in the university. Then the internal control evaluation index system can be successfully implemented. The reason why achieves the excellent level is closely related to the innovation of scientific research funds management in this university. As follows:

A. *Improve the internal control environment*
Z University leaders have strengthened the internal control construction in universities and strived to be the model for the implementation of the policy of "putting on the management service" to provide the exact guarantee for the internal control environment of scientific research funds. They have used the internal network platform of universities to train scientific research related personnel so that they can clearly realize the importance, seriousness and discipline of the internal control work of scientific research funds to improve their risk awareness of the use and management of scientific research funds, and to establish the main consciousness of internal control responsibility of the whole school. From the leaders to the staff of the scientific research departments, they all attach great importance to the construction of the internal control system for the scientific research funds within their own responsibilities. The division of labor clearly creates an excellent environment for the internal control of the scientific research funds of Z University. Moreover, relying on the working mechanism of all faculty members participating in internal control and actively adopting a continuous and effective dynamic cycle of internal control, the strategic goal of internal control of scientific research funds in Z University has been achieved. At the same time, an internal control organization for scientific research funds independent of scientific research, finance, assets and auditing departments in colleges and universities has also been established to organize, implement, supervise and evaluate the internal control construction of scientific research work, which really provides strong safeguard measures for the internal control of scientific research funds in colleges and universities from the administrative organization.

B. *Assess the funding risks dynamically*
The financial department of Z University conscientiously implements the "release" standards of the right to use funds, management of meetings and travelling, and scientific instruments and equipment. It takes the business process as the main line of internal control to check and fill gaps in the existing system, and to identify incompatible posts, and to carry out dynamic risk assessment form an efficient dynamic complete internal control system for firmly placing

the use of scientific research funds in the internal control horizon. Taking each management process of scientific research funds as a risk point, implement dynamic internal control at each node of budget preparation, actual expenditure and use performance, and controlling the whole process of scientific research activities improve internal supervision at key risk points to ensure the healthy development of scientific research activities.

C. *Strengthen the internal control activities*

Z University has revised a large number of documents related to the management of various scientific research funds since 2015. What is more important is to standardize and project the internal control management of scientific research funds and incorporate it into the overall internal control workflow of the internal management of the university, and continuously to improve it in the implementation. The management of Z uses standardized process and internal control system to provide guarantee for "management", revising internal control system, and sets up internal control team, compile internal control manual to achieve seamless connection of business. Ensuring responsibilities and fully sorting out the process to increase the intensity of budget tracking feedback will implement the regular meeting system for analyzing the effectiveness of fund management to improve the incentive mechanism and the responsibility recovery mechanism, and to coordinate with the financial and asset departments to "manage", which will realize asset sharing and improve the utilization rate of instruments.

D. *Ensure smooth communication of information*

Z University began to integrate the information platform in 2016, embedding the internal control work flow, measures, methods and standards of scientific research funds, using the database to carry out comprehensive dynamic analysis, realizing to share the information among departments, carrying out real-time statistics, monitoring and early warning on risk points of scientific research funds management, and using the analysis and monitoring functions of the scientific research funds information platform to promote the automatic completion of internal control evaluation indexes of scientific research funds in colleges and universities to improve internal control efficiency and enter a virtuous circle of active service. To create a good scientific research atmosphere for scientific researchers, the information platform is used to provide considerate convenience for "service", so as to improve the efficiency of the use of funds, to make online office appointments, to make use of Z University WeChat Public Platform and official website to disclose the progress in the use of scientific research funds in a timely manner.

E. *Project full-process audit supervision*

The leaders of Z University have begun to set up the specialized internal control organizations since 2016, and have formulated a normal internal control plan based on the annual funding for scientific research projects. They have carried out planned and step-by-step rectification of problems, such as the blind spots in internal control design and the implementation found in internal control assessment. In particular, the construction and operation of the internal control process, system and control subjects of scientific research funds in colleges and universities should be inspected and adjusted. With the continuous changes of internal and external factors, continuous supervision should be carried out in a cyclic manner. The evaluation should be based on a dynamic basic guarantee mechanism to realize the supervision of internal control effect. Strengthening the construction of audit supervision system in colleges and universities is another internal control method for Z. It is necessary to establish an all-round risk prevention and control system with internal control as the means and external supervision as the guarantee, so as to implement "management". The audit department should carry out special supervision over the project application, contract signing, acceptance results and the project funds currently which are implemented by the finance department, and also should pay more attention to the overall supervision during the whole event.

References

1. COSO, *Internal Control-Integrated Framework* (AICPA, 1992)
2. D. John, On collaborative governance-corporate social responsibility initiative working paper No. 2 (John F. Kennedy School of Government, Harvard University, Cambridge, MA, 2004)
3. D.A. Robertson, Agent-based models of a banking network as an example of a turbulent environment: the deliberate versus emergent strategy debate revisited. J. Complex. Issues Org. Manag., p. 56–71 (2003)
4. H. Haken, Ling Fuhua Translation, *Synergy: The Mystery of Nature* (Shanghai Translation Publishing House 1995 edition), p. 239
5. Commission on Global Governance, *Our Global Neighborhood* (Oxford University Press, Oxford, 1995), p. 2
6. Chris Ansell, Alison Gash, Collaborative governance in theory and practice. J. Public Adm. Res. Theor. **18**(4), 543–571 (2007)
7. P. Tian, Collaborative Governance Theoretical Research Framework and Analytical Model. Shanghai Jiaotong University Doctoral Dissertation (2013)
8. K. Emerson, T. Nabatchi, *Collaborative Governance Regime* (Georgetown University Press, Washington, DC, 2015), pp. 26–27
9. J. Silbernagel, K. Nixon, Collaborative scenario modeling reveals potential advantages of blending strategies to achieve conservation goals in a working forest landscape. Landsc. Ecol. **31**(5), 1093–1115 (2016)
10. J. Huang, Research on Construction and Application of Internal Control System of Scientific Research Funds in Public Universities under New Situation. Capital University of Economics and Business, Master's Degree Thesis (2017)
11. T.L. Saaty, *The Analytic Hierarchy Process: Planing, Priority Setting, Resource Allocation* (Pittsburgh, PA, 1990)

Multiobjective Optimization of Production Line Supply Based on Maximum Endurance Time and Rest Allowance

Sezen Korkulu and Krisztian Bóna

Abstract The endurance time of a muscle is the time which muscle sustain repeated contractions and it is one of key parameter for estimation of relaxation allowance and rest time. This paper extend and develop a new inventory model based on relaxation allowance and maximum endurance time. The model considered the maximum endurance time to calculate the rest time which necessary to prevent work-related back disorders for different quantities of handled items and item weights. The model analyzed with numerical example and different parameter sets to present validity under different scenarios. Finally, the analysis results were shown that the developed method using the rest allowance concept with maximum endurance time is suitable to integrate ergonomic aspects into a formal inventory cost model.

Keywords Endurance time · Rest allowance · Inventory management · Manual material handling · Work-related back disorders

1 Introduction

The work related musculoskeletal disorders and work-related back disorders are the most common occupational disorders which causes of lost or restricted work time and it is crucial to prevent WRMDs for promoting both economic and social sustainability. According to European agency for safety and health at work last report, work related musculoskeletal disorders one of the most crucial reason of the death in EU-28 proportion (%14.66) and the World proportion (%14.96) [11]. According to OSHA report; 30% of European workers suffer from back pain which is the most common work-related musculoskeletal disorders [10]. The maximum endurance time had been a key parameter for estimation of recovery times and quantifying muscular fatigue

S. Korkulu (✉) · K. Bóna
Depertment of Material Handling and Logistics Systems, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: sezen.korkulu@logisztika.bme.hu

K. Bóna
e-mail: krisztian.bona@logisztika.bme.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_60

to prevent work related musculoskeletal disorders. The endurance time of muscle force is the maximum time of holding force until the pain or fatigue occurs [9]. It has been investigated by a number of researchers that the relaxation allowance or rest time allows the recovery from muscular fatigue can be estimated from the endurance time of a muscular force [8, 18–22]. The constituted rest time and time standards were provided with an opportunity to increase the productivity of the workforce and lower the cost of a company. Consequently, the correct establishment of rest time, safe work design and precise rest allowance for muscular fatigue have to be considered in cycle times. In the literature, there can be found the great number of general and specific muscle group endurance time studies. Rohmert developed an exponential model for endurance time and endurance limit was 15% fMVC which means that forces till 15% fMVC can be held without tiring [19]. It was a general model and valid for all muscle groups and not dependent of a worker or task parameters. Rohmert developed endurance time models for specific muscle groups (shoulder, elbow, back and hand) [21]. Corlett and Manenica have been investigated the relationship between endurance time and relative force for the static pull and static torque [8]. Rose et al. investigated endurance time, pain and resumption time for fully flexed postures [22]. They found that endurance times in fully flexed postures differ little from those in more common postures. Therefore, the results showed that fully flexed postures may be assessed by more general prediction models for endurance. Garg et al. studied endurance times for different shoulder postures and their model does not have the endurance limit [12]. In the literature, there are few numbers of studies found that considers both ergonomics and inventory management [15]. Battini et al. have been investigated and analyzed the relationship between ergonomics and assembly system design techniques [2]. Battini et al. have been developed a new measurement technique with consideration of energy expenditure equations of Garg et al. [13] and rest allowance formulation of Rohmert [20] to simplifying the ergonomics assessment of each assembly task [4]. Battini et al. have been developed functions which are considered warehouse picking activities with the human availability and the rest allowance [3]. Battini et al. have developed a new multi-objective model for assembly line balancing which includes energy expenditure rate based on Predetermined Motion Energy System (PMES) [6]. Battini et al. have been developed a mixed-integer model which integrates assembly line balancing and parts feeding with incorporation ergonomic aspects based on rest allowances formulas of Garg et al. [13] and Price [5, 17]. Andriolo et al. have developed a lot-sizing model that considers multiobjective optimization of ergonomic aspects based on lifting index (LI) [1]. Battini et al. have been developed a mathematical model which investigates ergonomic lot size which is integrated Price [17] rest allowance formulation to the lot sizing model [7]. They have been investigated picking and storing motions with the energy expenditure rate for rest time assessment and did not consider the carrying motion and the maximum endurance time which has been a key parameter for better ergonomic assessment and calculation of rest time. Our investigation on the literature has shown that there is a few study which integrates ergonomics and inventory management. Therefore, there is a gap regarding maximum endurance time as an ergonomic measure and motions such as pushing, pulling, carrying and so on. The aims of the paper at hand is to extend and develop a new model for production line supply that covers rest allowance, carrying motion and maximum endurance time for preventing work-related musculoskeletal back disorders and ergonomic risks and to find the multi objective optimum which supports an economic and social sustainability. The rest allowance prevents disorders associated to manual material handling works in production supply process that includes picking, storing and carrying motions. The model developed in this paper helps to determine the optimal ergonomic total cost of production line supply and help to reduce ergonomic risks associated with repetitive work pace and help to increase the overall productivity.

2 Problem Description

A. *Framework of the handling activities and symbols*

The problem and framework studied in this paper is a single operator-single material model which covers placement of fixed amount of raw materials boxes from supermarket to the cart for production line supply process. The study examine preperation of production supply and does not cover transportation of products to the workstation. To promote manual handling of raw materials and to protect the worker from work-related musculoskeletal back disorders risks during handling, the new model investigated specific motions such as picking, storing and carrying where repetitive movement could employ ergonomic risks. As given in Fig. 1, the

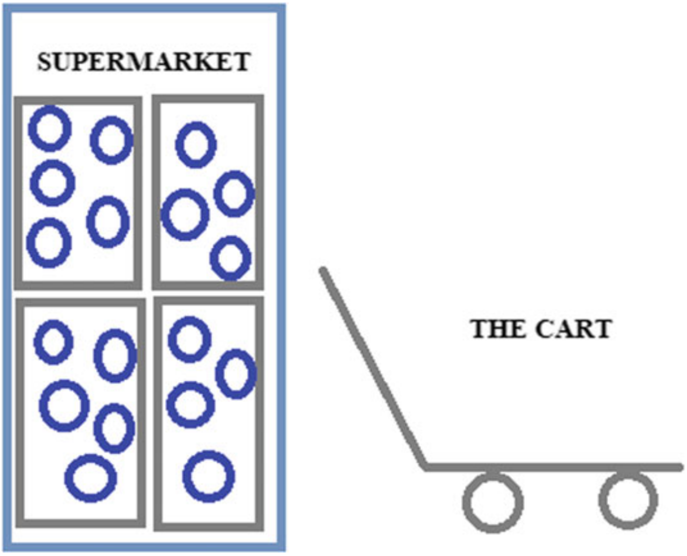


Fig. 1 The job cycle

job cycle consisted of; picking products from supermarket, carrying the products to the cart and placing the products to the cart.

The following assumptions will be made for developing the new model: the model considered a single worker and the model considered a single type of raw material.

The list of symbols and notations used in this paper:

$T_R(q)$	Rest time for avoiding ergonomic risks and disorders occur from fatigue (min)
$T_E(q)$	Endurance time for handling the lot q (min)
t_a	Availability time of the stock on the workstation (min)
w_u	Weight of the one unit (kg)
$fMVC$	Relative force (%)
$C(q)$	Total cost function ($)
$C_p(q)$	Total cost of picking ($)
$C_s(q)$	Total cost of storing ($)
$C_c(q)$	Total cost of carrying ($)
$C_I(q)$	Total cost of inventory (q)
$C_O(q)$	Total cost of operations ($)
$C_{TR}(q)$	Total cost of rest time ($)
$t_p(q)$	Unit picking time of a box from the supermarket (min)
$t_s(q)$	Unit storing time to the cart (min)
$t_c(q)$	Unit carrying time from the supermarket to the cart (min)
c_w	Unit worker wage ($/h)
c_h	Inventory holding cost ($/pcs/h)
q	Lot size (pcs)
$[Q/q]$	Number of cycle needed for handling total (Q) amount of items
T_{Rp}	Rest time needed for picking (min)
T_{Rs}	Rest time needed for storing (min)
T_{Rc}	Rest time needed for carrying (min).

3 Mathematical Model

A. *Rest allowance and maximum endurance time*

We used the Rohmert [19] rest time model where the rest time formulation included endurance time which is the maximum time during static load can be maintained. Based on the formulation developed by Rohmert [19, 20], the relaxation allowance and endurance time can be defined for a specific lot size as:

$$\mathrm{T_R(q)} = 1800 * (\mathrm{t/T_E})^{1.4} * (\mathrm{fMVC} - 0.15)^{0.5} \tag{1}$$

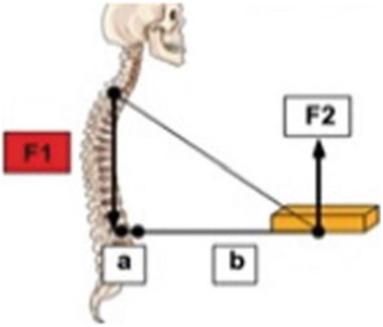


Fig. 2 The load on the spine

$$T_E(q) = -1.5 + (2.1/fMVC) - \left(0.6/fMVC^2\right) + \left(0.1/fMVC^3\right) \quad (2)$$

where, T_R is a rest allowance (% of total contraction time); q is a lot size; t is a contraction duration (working period of picking, storing and carrying) in a second; t/T_E is relative contraction duration; T_E is the endurance time; fMVC = fload/MVC is the relative force. As given in Fig. 2, the balance of the body can be determined by the force load on the spine.

To calculate the force load of picking, storing and carrying, we applied $F_{load} = (f * g * b)/a$, where f is a kg value of q; b is a distance between body axis and load axis; g is a gravitational acceleration, and a is a length of the load on the spine. MVC is equals F_{max} and, $F_{max} = (f_{max} * g * b)/a$ for picking, storing and carrying.

According to relaxation formulation, rest time is necessary if fMVC higher than 0.15, which leads to fMVC-0.15 ≥ 0. Therefore, fMVC of the picking and storing motions must be higher than 0.15. With this constraint, "T_R" can be expressed by the linear function of the tangent plane approximation as:

$$T_R = A_{re} * t + (B_{re} * fMVC) - C_{re} \quad (3)$$

where, A_{re} is the coefficient of contraction time, B_{re} is the coefficient of relative force and C_{re} is the constant of rest time. The linear function of rest time and endurance time can be defined as:

$$T_R = 3.641998153 * t + (32.22511961 * fMVC) - 5.322510481 \quad (4)$$

B. *The mathematical model*

We define the total cost function which include relaxation allowance cost as follows Eq. (5):

$$C(q) = C_I(q) + C_O(q) + C_{TR}(q) \quad (5)$$

where C(q) is the total cost of the production line supply, $C_I(q)$ is the total cost of inventory, $C_O(q)$ is the total cost of logistics operations and $C_{TR}(q)$ is the cost of relaxation allowance.

We define and extend [4], [7] the picking, storing, carrying, rest allowance and inventory cost functions as follows:

$$C_P(q) = t_p(q) * c_w * \left[Q/q\right] \tag{6}$$

$$C_s(q) = t_s(q) * c_w * \left[Q/q\right] \tag{7}$$

$$C_c(q) = t_c(q) * c_w * \left[Q/q\right] \tag{8}$$

$$C_I(q) = \left(t_a * q * \left[Q/q\right]\right) * c_h/2 \tag{9}$$

$$C_{TR}(q) = (T_{Rp} * t_p(q) + T_{Rs} * t_s(q) + T_{Rc} * t_c(q)) * c_w * \left[Q/q\right] \tag{10}$$

Equations (6)–(8) are the cost functions for picking, storing and carrying motions which represented the total time spending for picking, storing and carrying the total amount of items and also multiplied with unit worker wage as a cost of the worker. Equation (10) includes rest time for three activities to prevent any disorders occur from fatigue and it is also the cost function for rest time which is the total time spend on rest time for handling the total amount of items multiplied with unit worker wage. Equation (9) is the cost function of inventory holding. The total cost function can be rewritten as:

$$\begin{aligned} C(q) = {} & \left(t_a * q * \left[Q/q\right]\right) * c_h/2 \\ & + \left[t_p(q) * \left(T_{Rp} + 1\right) + t_s(q) * (T_{Rs} + 1) + t_c(q) * (T_{Rc} + 1)\right] * c_w * \left[Q/q\right] \end{aligned} \tag{11}$$

4 Results

To investigate the effects of rest time as a cost, we applied a numerical example, where total number of items equal to 2000 pcs, maximum voluntary contraction (F_{max}) for picking, storing abd carrying equal to 2452.5 N and it is calculated according to ISO Standard Ergonomics-Manual handling 11228 Part 1: Lifting and carrying limit for two-handed lifting which was 25 kg [14]. Time need of mounting 0.28 min/pcs, unit picking time of a box from the supermarket is 0.45 min, unit storing time to the cart is 0.45 min, unit carrying time from the supermarket to the cart is 0.75 min, unit worker wage is 18 \$/h, inventory holding cost is 0.1 \$/pcs/h, weight of each item is

1.5 kg. Anthropometrical parameters calculated as a and c is 5 cm for picking, storing and carrying, b is 50 cm for picking and storing, and carrying according to work related musculoskeletal disorders journal of Ministry of labor and social security of Turkey [16]. We simulated our model to find optimal total cost with fMVC $\geq$ 0.15. The minimum total cost equals $C(q) = 357.56\$$ and $C_{TR}(q) = 23.65\$$. We calculated the total cost of maximum lifting and carrying limit (25 kg) according to ISO Standard Ergonomics-Manual handling 11228 Part 1: Lifting and carrying limit for two-handed lifting where each lot size kg value equals 25 kg for production supply process under study and the total cost equals $C(q) = 550.82\$$ and $C_{TR}(q) = 203.39\$$. The saving of our model from using ergonomic rest time and maximum endurance time with comparison of the total cost without ergonomics equals 35.08%.

We calculated the savings obtained by using our new approach with different w_u (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9 and 3 kg) and savings are calculated with compared the cost of maximum lifting and carrying limit (25 kg) and minimum cost of the our model where fMVC $\geq$ 0.15.

As given in Fig. 3, an increase in unit weight increase the total cost and decreased the savings. Increase in unit weight will increase the relative force of the muscle for handling. Therefore, handling of heavier items will increase the rest time needs of worker and it leads higher rest time cost. Our model suggest to decrease lot size for heavier items to optimize rest time cost and ergonomic risk.

We calculated the optimal lot sizes with simulation of our new model to find both optimal lot size and total cost where fMVC $\geq$ 0.15. As given in Fig. 4, the increase in unit weight of item reduce the optimal lot size value. The reason of that the relative force (fMVC) is calculated according to unit item weight, therefore, increase in weight value decrease the lot size for reducing ergonomic risks.

Fig. 3 Savings for different weight of material

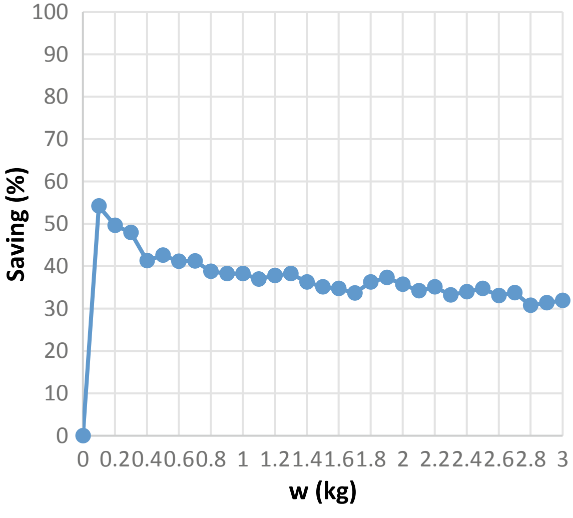
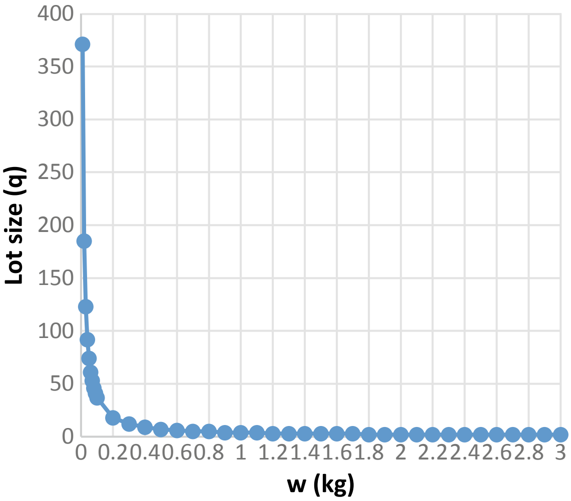


Fig. 4 Optimal lot size (q) in different weights of each item



We analyzed the new approach under different parameter values for determination of how total, operational, inventory, relaxation time cost changes. We applied three different item weight which is 0.2, 0.5 and 1 kg, unit storing time is changes between 0.06 and 0.30 min, unit picking time is changing between 0.06 and 0.30 min, unit carrying time is equal to 0.10–0.50 min, and other parameters were constant.

Results of the analysis were given in Figs. 5, 6, and 7 which illustrates that an increase in weight of the item will increase especially the total cost, cost of rest time and operational cost, decrease the lot size. In particular, there is no big difference for the inventory cost when the weight of the item changes. As can be seen, the optimal solutions which model suggests that lower the weight of item or decrease the lot size for a decrease in the total cost and the rest time cost.

5 Conclusion

The work-related musculoskeletal disorders are the most common health problem in the World. Especially, manual handling, uncomfortable working positions and repetitive movements, all actions which are often associated with working in a manual job are the most common causes of musculoskeletal disorders. Although many research has contributed on the lot sizing, especially work related musculoskeletal back disorders and ergonomics in lot sizing has received very little attention in the literature so far. In this paper, to improve ergonomic conditions and reducing the work related musculoskeletal back disorders risk, the ergonomics aspects were integrated into a new extended cost model. There is a lack of investigation which integrates maximum endurance time as an ergonomic measure which is a key parameter to find rest allowance need. The new model helps to decrease the ergonomic risk of most common

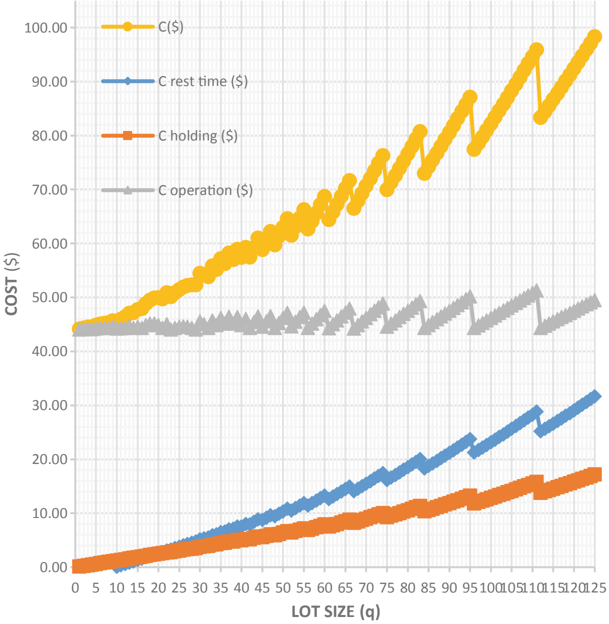


Fig. 5 Total cost curve and rest time cost curves for alternative lot-size where unit weight 0.2 kg

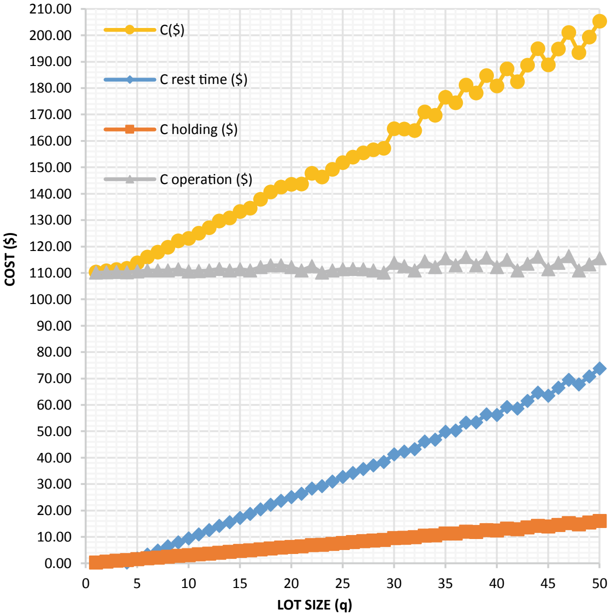


Fig. 6 Total cost curve and rest time cost curves for alternative lot-size where unit weight 0.5 kg

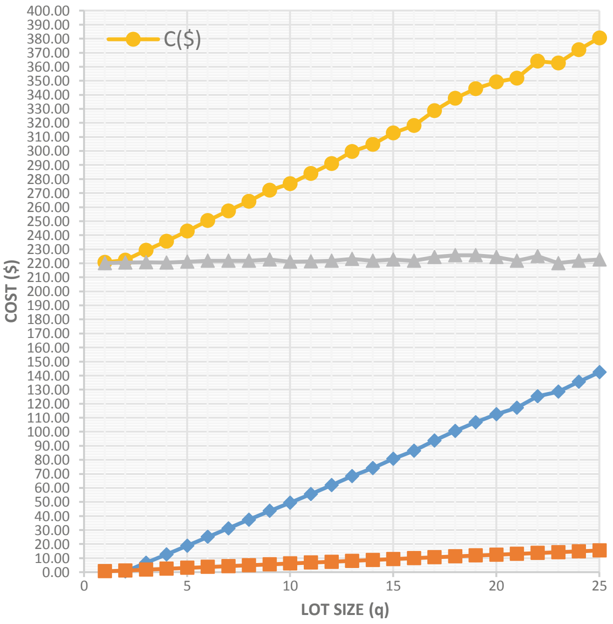


Fig. 7 Total cost curve and rest time cost curves for alternative lot-size where unit weight 1 kg

motions in manual material handling that are picking, storing and carrying motions. The developed model analysis results were shown that the developed method using the rest allowance concept with maximum endurance time is suitable to integrate ergonomic aspects into a formal inventory cost model. The integrated model can help to reduce the ergonomic risks regarding to musculoskeletal back disorders resulting from repetitive manual handling work and improve ergonomics aspect in industrial practices and promote both economic and social sustainability.

The future work could investigate first, a situation with handling more raw materials as multi material—single operator and multi material—multi operator for better understanding of the relationship between lot size and the operator and how to improve the ergonomic conditions and social sustainability in industrial practices. Furthermore, investigation in the handling of materials in semi-automated production line where the repetitive motion and force of movement could employ the work related musculoskeletal disorders and ergonomic risks, would be valuable extension of these research.

References

1. A. Andriolo, D. Battini, A. Persona, F. Sgarbossa, A new bi-objective approach for including ergonomic principles into EOQ model. Int. J. Prod. Res. **54**(9), 2610–2627 (2016). https://doi.

org/10.1080/00207543.2015.1113324

2. D. Battini, M. Faccio, A. Persona, F. Sgarbossa, New methodological framework to improve productivity and ergonomics in assembly system design. Int. J. Ind. Ergon. **41**(1), 30–42 (2011). https://doi.org/10.1016/j.ergon.2010.12.001
3. D. Battini, M. Calzavara, A. Persona, F. Sgarbossa, Linking human availability and ergonomics parameters in order-picking systems. IFAC-PapersOnLine **48**(3), 345–350 (2015). https://doi.org/10.1016/j.ifacol.2015.06.105
4. D. Battini, C.H. Glock, E.H. Grosse, A. Persona, F. Sgarbossa, Ergo-lot-sizing: considering ergonomics in lot-sizing decisions. IFAC-PapersOnline **48**(3), 326–331 (2015). https://doi.org/10.1016/j.ifacol.2015.06.102
5. D. Battini, M. Calzavara, A. Otto, F. Sgarbossa, The integrated assembly line balancing and parts feeding problem with ergonomics considerations. IFAC-PapersOnLine **49**(12), 191–196 (2016). https://doi.org/10.1016/j.ifacol.2016.07.594
6. D. Battini, X. Delorme, A. Dolgui, A. Persona, F. Sgarbossa, Ergonomics in assembly line balancing based on energy expenditure: a multi-objective model. Int. J. Prod. Res. **54**(3), 824–845 (2016). https://doi.org/10.1080/00207543.2015.1074299
7. D. Battinia, H.C. Glock, E.H. Grosse, A. Persona, F. Sgarbossa, Ergo-lot-sizing: an approach to integrate ergonomic and economic objectives in manual materials handling. Int. J. Prod. Econ. **185**, 230–239 (2017). https://doi.org/10.1016/j.ijpe.2017.01.010
8. E.N. Corlett, I. Manenica, The effects and measurement of working postures. Appl. Ergon. **11**(10), 7–16 (1980). https://doi.org/10.1016/0003-6870(80)90115-5
9. M. Eksioglu, Endurance time of grip-force as a function of grip-span, posture and anthropometric variables. Int. J. Ind. Ergon. **41**, 401–409 (2011). https://doi.org/10.1016/j.ergon.2011.05.006
10. EU-OSHA, Research on Work-Related Low Back Disorders (2000). Available at: https://osha.europa.eu/en/tools-and-publications/publications/reports/204
11. EU-OSHA, An international comparison of the cost of work-related accidents and illnesses (2017). [Online]. Available: https://osha.europa.eu/en/tools-and-publications/publications/international-comparison-cost-work-related-accidents-and/view
12. A. Garg, K.T. Hegmann, B.J. Schwoerer, J.M. Kapellusch, The effect of maximum voluntary contraction on endurance times for the shoulder girdle. Int. J. Ind. Ergon. **30**, 103–111 (2002). https://doi.org/10.1016/S0169-8141(02)00078-1
13. A. Garg, D.B. Chaffin, D.H. Gary, Prediction of metabolic rates for manual materials handling jobs. Am. Ind. Hygiene Assoc. J. **39**(8), 661–674 (1978). https://doi.org/10.1080/0002889778507831
14. ISO 11228-1, Ergonomics—Manual handling—Part 1: Lifting and carrying (2003). [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:11228:-1:ed-1:v1:en
15. S. Korkulu, K. Bóna, Ergonomics as a social component of sustainable lot-sizing: a review. Period. Polytech. Soc. Manag. Sci. **27**(1), 1–8 (2019). https://doi.org/10.3311/PPso.12286
16. Ministry of Labor and Social Security of Turkey, Work related musculoskeletal disorders. Occup. Health Saf. J. (2007)
17. A.D. Price, Calculating relaxation allowances for construction operatives—part 1: metabolic cost. Appl. Ergon. **21**(4), 311–317 (1990). https://doi.org/10.1016/0003-6870(90)90202-9
18. W. Rohmert, Ermittlung von Erholungspausen für Statische Arbeit des Menschen. Internationale Zetischrift fur Angewandte Physiologie Einschliesslich Arbeitphysiologie **18**, 123–164 (1960). https://doi.org/10.1007/BF00698869
19. W. Rohmert, Problems in determining rest allowances. Part 1: use of modern methods to evaluate stress and strain in static muscular work. Appl. Ergon. **4**(2), 91–95 (1973). https://doi.org/10.1016/0003-6870(73)90082
20. W. Rohmert, Problems in determining rest allowances. Part 2: determining rest allowances in different tasks. Appl. Ergon. **4**(3), 158–162 (1973). https://doi.org/10.1016/0003-6870(73)90166-X
21. W. Rohmert, M. Wangenheim, J. Mainzer, P. Zipp, W. Lesser, A study stressing the need for a static postural force model for work analysis. Ergonomics **29**(10), 1235–1249 (1986). https://doi.org/10.1080/00140138608967237

22. L. Rose, R. Ortengren, M. Ericsson, Endurance, pain in fully flexed postures. Appl. Ergon. **32**, 501–508 (2001). https://doi.org/10.1016/S0003-6870(01)00016-3

Study on Site Selection Evaluation of Police Drone for the Disposal of Abnormal Moving Vehicles

Zhaowei Ding, Xin Wang, Lei Wang, Jianhua Yang, and Ai Wang

Abstract Abnormal moving vehicles pose a greater threat to people, vehicles and objects around them. Compared with the traditional way of sending out police, the police drone (UAV) is faster and more agile, can effectively detect and stop them. In order to improve the disposal efficiency of police drone, this paper takes into account the environmental risk road resilience and police defense level. Site selection evaluation is carried out. As example a police station, combined with GIS technology, quantitative and qualitative evaluation to verify the feasibility of the site selection evaluation scheme is used. This method not only considers the potential risk distribution of the region, but also adds the influence of police resources on site selection evaluation. The relationship between regional risk assessment and scientific site selection is established, to provide better decision support for the site selection layout of police drone for disposing abnormal vehicles.

Keywords Abnormal moving vehicles · Unmanned aerial vehicle · Site assessment · Geographical information system

Z. Ding · J. Yang · A. Wang
University of Science & Technology Beijing, Donlinks School of Economics and Management, Beijing, China
e-mail: dingzhaowei1013@126.com

J. Yang
e-mail: yangjh@ustb.edu.cn

A. Wang
e-mail: wangai22222@126.com

X. Wang (✉) · L. Wang
Criminal Investigation Police, University of China, Shenyang, China
e-mail: leonwang521@126.com

L. Wang
e-mail: leonwang521@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_61

1 Introduction

Moving vehicles have attributes about dynamic and autonomous decision-making, to which the interception success rate is low, and the interceptors are in great danger. In case of abnormal driving conditions, it is necessary to take alert and action immediately. In particular, under the constant threat of international terrorist organizations, the explosive terrorist incidents caused by vehicle-mounted bombs and drive-by collisions pose a great threat to government management and people's lives [1, 2]. Such examples relay how ubiquitous and attractive highly visible transportation-oriented attacks are for modern terrorist operatives [3]. However, due to the impact of traffic conditions and the uncertainty of abnormal vehicle information, the best decision is to send air support forces to detect and intercept while sending the police to participate in the disposal. Because of its small size, convenient delivery, and multi-point reserve in urban areas, UAVs are an important resource and equipment for future police emergency handling. In particular, the investigation, tracking and stopping of vehicles on the road has obvious advantages.

The UAV can not only undertake on-site monitoring, reconnaissance and hunting, traffic patrol, large-scale event security, search and rescue, but also participate in tracking and interception, which can save police force and reduce the risk of police officers. Pablo Garcia-Aunon et al. took advantage of the convenience and rapidity of UAV aerial monitoring to develop a framework for information transmission visualization and emergency monitoring of smart city by UAV [4]. Zhulhua et al. studied the avoidance of obstacles during the flight of multi-UAVs, so as to maximize the flight surface [5]. With the characteristics of low cost, light weight, strong mobility and strong adaptability, Menglan Hu et al. studied the scheduling of UAVs and vehicle coordination [6]. But there are few researches on site selection evaluation. The risk assessment of the area in charge directly determines the location and investment quantity of police drone, which is the basis of carrying out UAV participation in police work.

Athanasios Ch. Kapoutsis et al. adopted the technique of region representation to divide the region into the same unit and then consider the configuration problem of multiple operation units [7]. This method can quickly find out the best strategy of the delivery point in the unit region when the service object of the region is not treated differently, but it ignores the difference of risks in the real environment, and tasks cannot be assigned uniformly in different regions. In terms of regional risk assessment, Masoud Khanmohamadi et al. use the zero-sum game to consider the population density and construction facilities around a railway, so as to assess the vulnerability of transport security of dangerous goods under terrorist attacks [8]. Ding used the target loss probability model to make decisions on the limited resource allocation of each target to prevent terrorist attacks [9]. The above scholars discussed the vulnerability of regions or routes in the context of hypothetical violent terrorist attacks, Corri Zoli et al. introduced game confrontation into the study [10]. The logic of the study is to invest in security and defense according to the game results. The influence of regional environment on police work is not included in the decision variables of input, so the evaluation of the police input does not have much reference value.

The first step in risk assessment is to identify the target of the assessment. We are more concerned about the hazards that can be caused by an erratically moving vehicle. Hu et al. supposed that intentional attack is different from natural disasters [11]. The intentional attacker selects the attack target and the attack strategy according to the situation of the city defense strategy and the type of city defense resources, thus showing the rationality of the attacker's decision. Therefore, a deliberate attack cannot be considered a simple random event. An analysis should be made of the nature of the attack and a risk assessment of the potential attack area based on the characteristics of the attack. In addition, police vehicles involved in the disposal are restricted by traffic conditions and road accessibility, they are easily detected or provoked. In addition, the existing sky-eye projects are mostly fixed-point directional road monitoring, and some camera devices with dynamic monitoring functions also have blind spots [12]. In response to the above problems, relying on the rapidity, dexterity, and safety features of police drones is an important means to compensate police officers for handling abnormally moving vehicles. The location of police drone directly determines its performance and the disposal of police situation. On the basis of UAV performance, scientific risk analysis and assessment of the management area and reasonable planning are the key points in the study of UAV site selection.

The article is divided into five aspects for research, the first part is the overview part; the second part is the further analysis of the problem and the selection of risk assessment indicators; the third part is the establishment of an index system and the determination of weight; the fourth part is the simulation of the real jurisdiction, and the fifth part is the conclusion.

2 Problem Description and Index Selection

Unlike the risk assessment of unintentional, natural disasters or static events, abnormally moving vehicles are a source of danger that starts in road traffic and radiates into the surrounding environment. At the Central Political and Law Working Conference on January 17, the emphasis was placed on the prevention and management mechanisms of road traffic and other areas, as well as the need for monitoring and early warning and emergency response mechanisms [13]. But so far, there has been no research on defining the vulnerability analysis of abnormally moving vehicles in the police management area to guide the allocation of police resources. Many qualitative methods for calculating risk follow this original equation: Risk = Threat × Vulnerability × Consequences.

According to the above formula, the probability of occurrence of T is a very difficult variable to evaluate, because the probability of dependence on external factors is less predictable. For this type of realistic problem, we adopt qualitative and quantitative research methods.

2.1 Selection Principles of Risk Assessment Indicators

When assessing the risk of being attacked in a particular scene or area, it is not possible to use historical data to estimate and judge, especially if such an area has never experienced such an attack, which does not mean that its own risk value is low. Therefore, we no longer start from the probability of the event, but consider the impact of regional environmental attributes on the judgment of intentional abnormal vehicle attacks, and the police's handling and response capabilities after an emergency. For the risk factors of abnormal moving vehicles, it will not only deter the safety of surrounding pedestrians, but also have a certain impact on the level of road traffic and the situation of construction units. Especially for dangerous drivers or terrorists who use driving as a means of crime or terrorist attack, the choice of the place of implementation is more targeted. In view of this, we divide regional risk factors into regional environmental vulnerability factors and regional defense level factors as shown in Fig. 1.

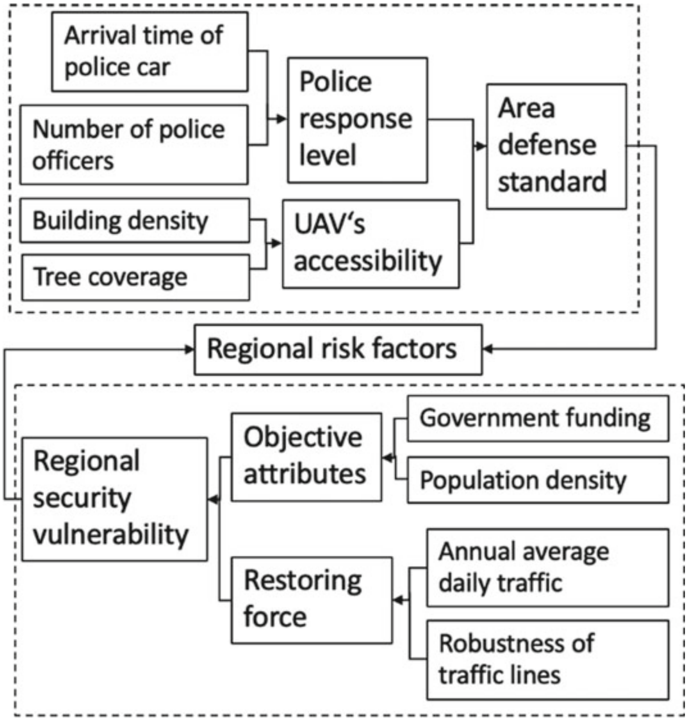


Fig. 1 Regional risk division

2.2 Vulnerability Factors of Regional Environment

Urban structure refers to the form and mode of mutual relationship and interaction among the various elements in a region, mainly including economic structure, social structure and spatial structure. According to this definition, according to the attributes of urban areas and the possible consequences of abnormal moving vehicles on urban areas, we consider regional objective attributes and regional resilience as the criteria to consider regional environmental vulnerability. See Fig. 2 for details.

1. *Calculation method of regional objective attribute*

It is difficult to predict the route and time of abnormal moving vehicles, so potential targets must be determined. Potential targets are places with some strategic elements. We assess the consequences in terms of the most serious characteristics of the event, that is, from an economic point of view and from a loss of life point of view. It allows the identification of key consequences for each asset and which asset needs to be risk managed. Specifically, it includes the following two aspects:

(a) *Regional government investment*

In order to determine the priority of each district within the jurisdiction, it is mainly based on the public data of public security expenditure in the annual Department accounts of the judicial department of the county and district government, in which only basic expenditure items are considered.

The budget data is used to prioritize infrastructure. Types of infrastructure include religious, commercial and government sites, transport hubs, parks, medical centers, museums, and educational and recreational sites. By dividing the total regional basic expenditure by the number of infrastructures in the region, an average security index of the region can be obtained. In this way, it is convenient to study the distribution of investment in different areas.

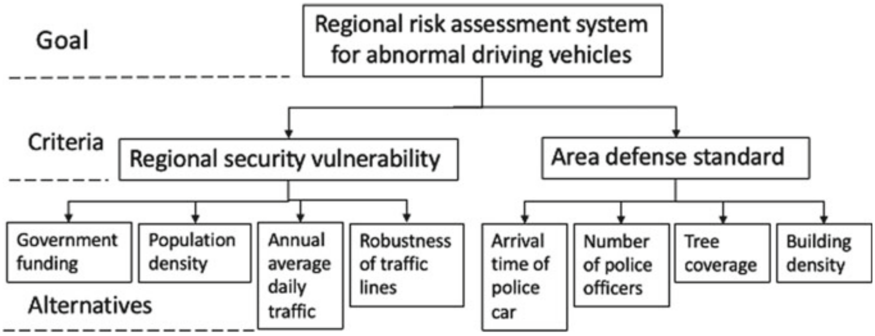

Goal
Regional risk assessment system for abnormal driving vehicles
Criteria
Regional security vulnerability
Area defense standard
Government funding
Population density
Annual average daily traffic
Robustness of traffic lines
Arrival time of police car
Number of police officers
Tree coverage
Building density
Alternatives

Fig. 2 Comprehensive evaluation system

(b) *Regional population density*

In addition to damaging the network and transportation infrastructure, the attacks may have targeted the number of people affected. Most attacks only target people. Such as the 2017 concert in Manchester, a football stadium in France in November 2015, a concert in the United States in 2017 or the 2017 Islamic consultative conference in Iran. In a worst-case scenario, a terrorist attack on a truck carrying chlorine gas kills 100 people every second and about 100,000 in a half an hour in a densely populated area, according to research conducted by the U.S. Army research laboratory [14, 15]. Therefore, in the study of risk accidents caused by abnormal moving vehicles, the exposure of people on both sides of the road is also worth considering. We used the demographic layer in the GIS software to determine the number of people exposed to the attack in the study area.

2. *Calculation method of regional restoring force*

The resilience of the region is mainly concerned with the accessibility and stability of road traffic in the region under the attack of abnormal moving vehicles. Specifically including the following two points.

(a) *Regional road traffic flow*

Traffic flow refers to the number of participants who actually pass through a certain point or section of the road in a unit of time. In the absence of special instructions, traffic volume refers to motor vehicle traffic, and refers to the number of vehicles per unit of time in both directions. It is the actual capacity of the road section measurement index, is also the main basis for road classification and determination of road grade.

In this study, the average traffic volume of the line is used. As it has a certain time attribute, we use average daily traffic flow (ADT). Since this kind of crisis occurs during the day, we only consider the traffic flow during the daytime. Formula is as follows:

$$Q_r = Q * A_r / R \tag{1}$$

where

A_r: daytime traffic coefficient, the daytime coefficient of the whole network is about 0.80

R: the number of hours in the day is 16

Q: 24-hour traffic flow in each forecast year.

(b) *Robustness of regional circuits*

The substitutability after line interruption is the robustness of the road traffic network. Considering the subjectivity of drivers, we take the crime hotspots within the jurisdiction as the sites with high probability to be attacked. The more crime hotspots within the jurisdiction, the higher the probability of criminal attacks and the worse the stability of the traffic network. In this study, roads within the jurisdiction will be simulated into a network diagram, and each node in the network represents the convergence point of multiple lines. The number of nodes directly connected to each other is called node degree. The average node degree of the network is the average of all node degrees in the region, which can reflect the connectivity of the regional network. By integrating the factors, the robustness of the regional circuit can be reflected by the following formula:

$$R = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}/nN \tag{2}$$

where

R: robustness of regional circuits

a_{ij}: the connectivity between node i and node j in the region, when point i and j have an edge, $a_{ij} = 1$; otherwise, $a_{ij} = 0$

N: the number of nodes in a region

n: the number of crime hotspots in the region.

2.3 Area Fortification Level

Once it is determined that the vehicle is running abnormally, The police station will send police cars to deal with it as soon as possible. However, police resources are limited by the accessibility of road traffic. When police cars are dispatched to unusual targets, police drones should be dispatched to investigate them as soon as possible. Therefore, in the assessment of the level of fortification, we consider the level of police disposal and the convenience of UAV participation to measure.

1. *Police response level*

The handling level of the police team to the abnormal driving vehicles is reflected in the police time, as well as the police and police equipment. The reference standard is the compensation coefficient of regional risk police force, CP. It can reflect the overall capacity of the police force to deal with the jurisdiction. It is characterized by quick response. The response time was calculated according to formula:

$$t_p = t_{p1} + \frac{s_p}{u_p} \tag{3}$$

where

t_p: the time of quick response (min)

Table 1 The value of regional risk compensation coefficient

The time of quick response	Emergency rescue radius	The CP values
10 min	s_1	0.70
20 min	s_2	0.90

Police response level (PRL) $= 1 - CP$

t_{p1}: police response time (min), according to the average police time in the evaluated area to determine

s_p: the average distance of police to the precinct (m)

u_p: the speed of a police car in a district (m/s).

With reference to the provisions of ICI Mond law on compensation for sudden disasters such as explosion and fire, and in combination with the rapid response mechanism of the police in China, the areas where a general alert can arrive within 10 min are covered by the police. For rush hour traffic and places far from the police station, the police time shall be 20 min. For the part outside the police force coverage area, CP is 1. According to Eq. (3), when the exit time of the evaluated area and the driving speed of the police car are known, the radius sf of the covered area can be determined, which is calculated according to Eq. (4). The value of regional risk compensation coefficient CP is shown in Table 1.

$$s_p = \left(t_p - t_{p1}\right)u_p \tag{4}$$

2. *The convenience of UAV disposal*

Compared with police cars, the most prominent feature of UAVs in driving convenience is that it can ignore the accessibility of road traffic, but it is also limited by the objective environment around the road, such as high-rise buildings and dense forest vegetation, which will affect its commuting efficiency. We used the statistical layer of tree coverage in GIS software to determine the tree coverage in the study area.

3 Index System and Index Weight

The relationship between indicators is complex, involving a wide range of fields. In the risk analysis, it is necessary to decompose the target index, clarify the relationship and weight of each index, and construct the UAV site selection index structure system for abnormal moving vehicles.

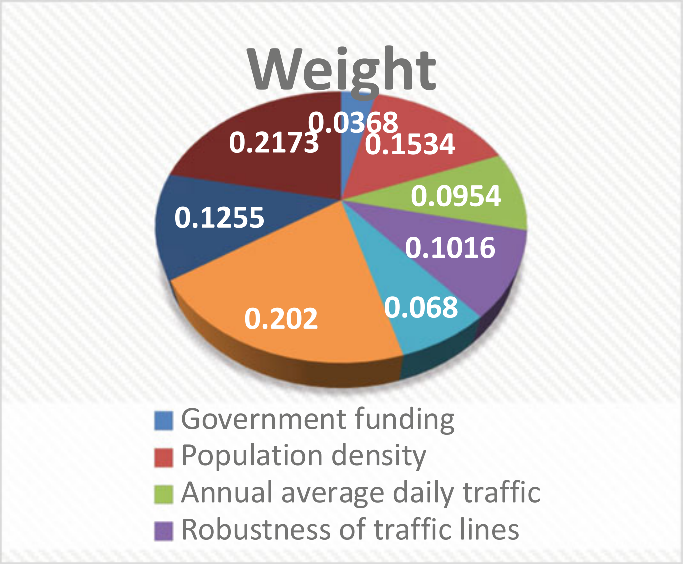


Fig. 3 Index weight diagram

3.1 Establishment of Evaluation Index System

- Target layer: that is, the set of target problems to be solved. The harm caused by abnormal moving vehicles to the surrounding area is taken as the target layer to reflect the risk.
- Criterion layer: that is, the set of factor indicators that constitute the target problem. In this index system, the risk caused by abnormal moving vehicles is affected by environmental vulnerability and environmental defense level, so they are taken as the criterion layer.
- Indicator layer: that is, the basic indicator set, which belongs to the indicator set of various factors in the criterion layer. The metrics in the metrics layer are the foundation for achieving the goal problem, including eight metrics (Fig. 3).

3.2 Determine the Weight of Evaluation Indicators

In this paper, the analytic hierarchy process (AHP) is used to determine the weight coefficient of the index. These factors are subordinate to each other at different levels, and the factors at the same level interact with each other to form a hierarchical structure system. In AHP, the target problem is divided into three layers: target layer, criterion layer and index layer, then a weight matrix is assigned and a weight coefficient is calculated.

Consistency check. Taking the n-order weight matrix as an example, the oneness index is expressed as follows:

$$C.I = \frac{\lambda_{\max} - n}{n - 1} \tag{5}$$

$\lambda_{\max}$ is the maximum characteristic root.

$$C.R = \frac{C.I}{R.I} \tag{6}$$

$C.R$ is the consistency ratio; $R.I$ is a random consistency indicator (please refer to the calculation table of random consistency indicator).

If $C.R$ is less than or equal to 0.1, the matrix satisfies the check condition. Otherwise, i and j scales need to be reassigned to verify the matrix. They are the rows and columns of index fuzzy matrix.

According to the calculation principle of analytic hierarchy process and the indicator system constructed in the previous section, the weight coefficients of the indicator layer and criterion layer determined in this indicator system are as follows.

4 Case Risk Assessment Study

This paper takes the area under the jurisdiction of a police station in Beijing as the research object, covering a total area of 17 km^2, involving 165 units, 22 universities and primary schools, covering a population of nearly 250,000. There are no chemical plants and dangerous materials in this area, there are more schools and shopping malls, the population density of this area is high, the area under its jurisdiction is large, and the social influence after the disaster is wide. According to the service regulations of police stations of Beijing municipal public security bureau, each police station is divided into four police working areas according to the geographical space characteristics, the total amount of service and public security situation of the area under its jurisdiction, and each working area has a police standby position (Fig. 4). Through the evaluation of this study, it can be determined that an optimal police standby position for UAV is in charge of the whole police station's UAV service.

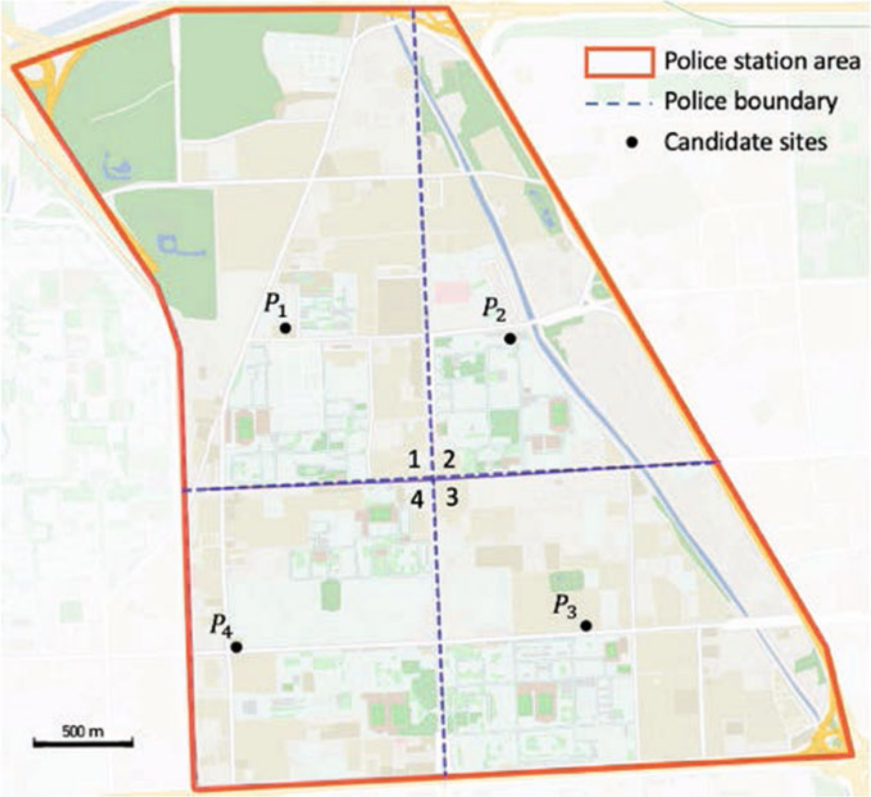


Fig. 4 The jurisdiction of the police station

4.1 Risk Factors Analysis of Abnormal Driving Vehicles in Each Police District

We set the police stations to be evaluated as No. 1, No. 2, No. 3 and No. 4 respectively. According to the evaluation system, the characteristics of the police stations will be analyzed in the order of the police districts:

1. *Regional vulnerability*

The data in Table 2 are shown as the formula of the matrix:

$$D_1 = \begin{bmatrix} 39.40 & 8100 & 6200 & 4.19 \\ 23.64 & 8300 & 6500 & 3.50 \\ 39.40 & 8700 & 6900 & 3.56 \\ 27.58 & 8900 & 7100 & 3.73 \end{bmatrix}$$

Table 2 Regional security vulnerability

Police working areas	Regional security vulnerability			
	Objective attributes		Restoring force	
	Government funding	Population density	Annual average daily traffic	Robustness of traffic lines
1	39.40	8100	6200	4.19
2	23.64	8300	6500	3.50
3	39.40	8700	6900	3.56
4	27.58	8900	7100	3.73

Table 3 Area defense standard

Police working areas	Area defense standard			
	Police response level		UAV's accessibility	
	Number of police officers	Arrival time of police car	Building density (%)	Tree coverage (%)
1	15	0.3	34.10	23
2	18	0.1	31	18
3	17	0.1	48.00	13
4	17	0.3	53.30	16

2. *Regional defense level*

The data in Table 3 are shown as the formula of the matrix:

$$D_2 = \begin{bmatrix} 15 & 0.3 & 34.1 & 23 \\ 18 & 0.1 & 31 & 18 \\ 17 & 0.1 & 48 & 13 \\ 17 & 0.3 & 53.3 & 16 \end{bmatrix}$$

3. *Vector modularization*

$$r_{ij} = \frac{d_{ij}}{\sqrt{\sum_{i=1}^{m} d_{ij}^2}} \tag{7}$$

Based on the weights obtained above, the following table can be obtained:

$$V = \left(v_{ij}\right)_{4\times 8} = \begin{bmatrix} 0.0218 & 0.0730 & 0.0442 & 0.0567 & 0.0304 & 0.1355 & 0.0502 & 0.1398 \\ 0.0131 & 0.0748 & 0.0464 & 0.0473 & 0.0365 & 0.0452 & 0.0456 & 0.1094 \\ 0.0218 & 0.0784 & 0.0492 & 0.0482 & 0.0344 & 0.0452 & 0.0707 & 0.0790 \\ 0.0153 & 0.0802 & 0.0507 & 0.0505 & 0.0344 & 0.1355 & 0.0785 & 0.0973 \end{bmatrix}$$

4.2 *The TOPSIS Method Calculates the Value of Regional* **Comprehensive Evaluation**

The normal solution is composed of the largest element of each column vector of V (the smallest element of the 4th, 5th, and 6th column vector).

$$v^{+}(0.0218, 0.0802, 0.0507, 0.0567, 0.0365, 0.1355, 0.0785, 0.1398)$$

$$v^{-}(0.0131, 0.0730, 0.0442, 0.0473, 0.0304, 0.0452, 0.0456, 0.0790)$$

The (Euclidean) distance between the scheme and the positive ideal solution:

$$S_i^{+} = \sqrt{\sum_{j=1}^{8}\left(v_{ij} - v_j^{+}\right)^2} \tag{8}$$

$$S_i^{+} = (0.0305, 0.3199, 0.1095, 0.0434)$$

$$S_i^{-} = \sqrt{\sum_{j=1}^{8}\left(v_{ij} - v_j^{-}\right)^2} \tag{9}$$

$$S_i^{-} = (0.1097, 0.0311, 0.0278, 0.0984)$$

The relative proximity between the scheme and the positive ideal solution is defined as:

$$C_i^{+} = \frac{S_i^{-}}{S_i^{+} + S_i^{-}} \tag{10}$$

$$C_i^{+} = (0.7824, 0.0886, 0.2024, 0.6939)$$

After normalization, we get:

According to the scores, we should choose the No. 1 area (Table 4).

Table 4 Police working areas point

Police working areas	Points
1	0.4427
2	0.0501
3	0.1145
4	0.3926

4.3 Suggestions on the Standby Point of Police UAV

After communication with the public security organs, the high-crime receiving and handling police in the area from 2017 to 2019 shall be taken as a reference, experts the way of marking to identify the key positions (the amount of the police in each police group, who is responsible for receiving crimes is 8), and 21 hot spots of crime shall be marked in the area (Fig. 5). According to the distribution of cases along the road by Beijing Public Security Traffic Management Bureau n, considering the police car's arrival time t and the distribution of closed-circuit surveillance m around it, $n \times t \times m/L$ is adopted to simulate the high-risk routes with hot spot distribution streets as samples L (Fig. 6). Based on the alternative points, the midpoint of each high-risk route is used as the target point for the UAV to patrol the route, and the flight path of the UAV is obtained (Fig. 7; Table 5).

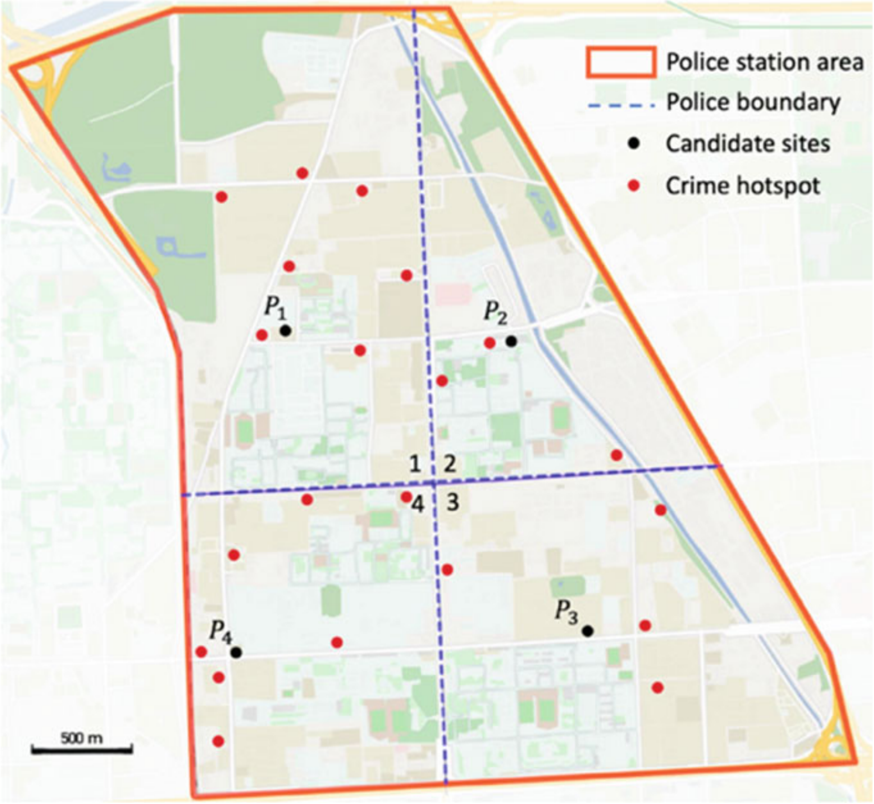


Fig. 5 Crime hot spots

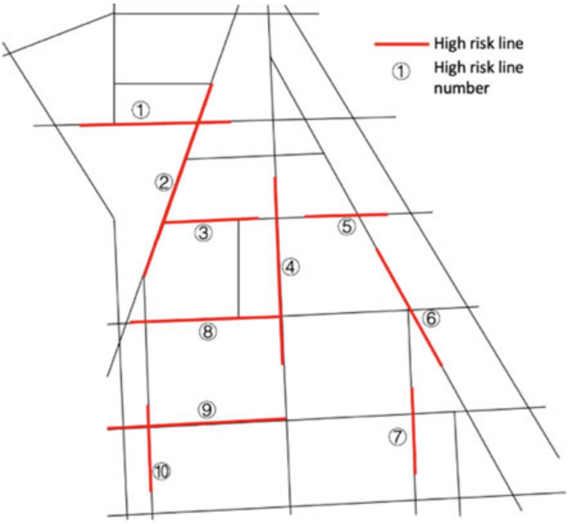


Fig. 6 High risk line

$$L_n = \sum_{i=1}^{10} l_{ni}; \quad n = (1, 2, 3, 4) \tag{11}$$

where

l_i: The flight distance from the place of arms to the center point of the high-risk line;

L_n: The total flight distance of the UAV from the places of arms n to the center of all routes.

As the places of arms are set up by each police station according to the security characteristics of its own jurisdiction, it takes into account the relevant characteristics of crime control and convenient service, etc. The calculated results are consistent with the risk assessment results (Table 6).

5 Conclusions and Prospects

Rooted in environmental criminology, this study links potential public safety problems in geographic environments to a police resource allocation. With the aggravation of the threat of abnormal driving vehicles to social security, the traditional management cannot control the occurrence of such crimes well, but the emergence of police drone can solve this problem. In order to ensure the efficient disposal of UAVs, scientific evaluation and selection of its site selection and standby points are required. This study starts from the vulnerability of the environment itself and the level of regional defense, and involving eight specific indicators, such as capital

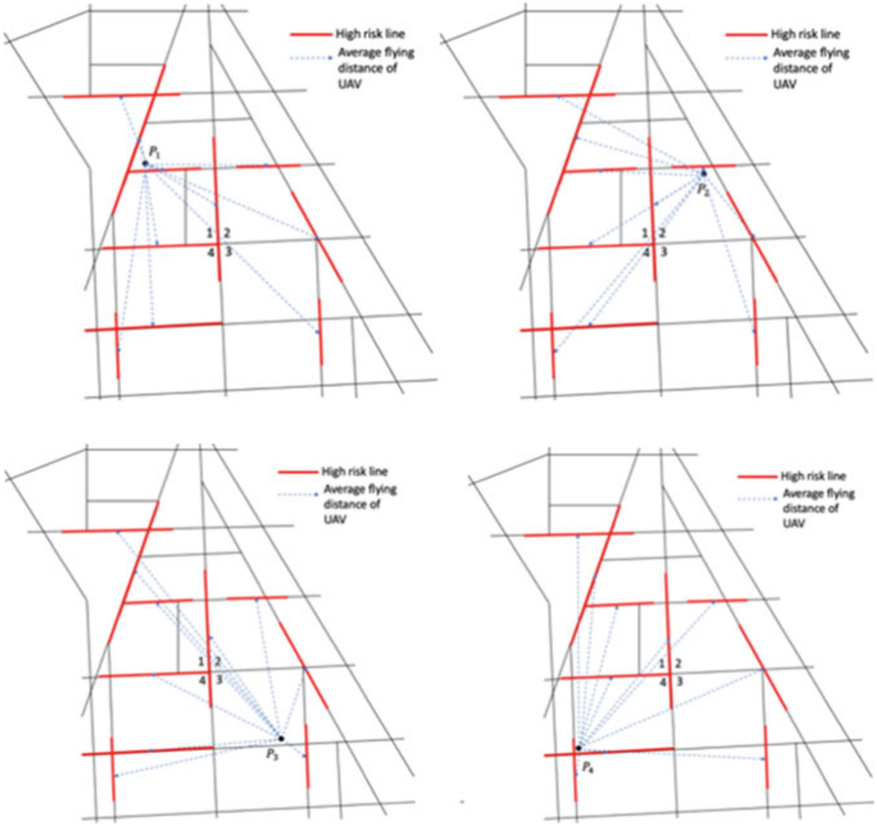


Fig. 7 UAV flight path

Table 5 Territorial disposal target

Place of arms	Number of crime hotspots	The length of high-risk lines	High risk line
p_1	6	4.59	①②③④⑧
p_2	3	1.62	④⑤⑥
p_3	4	1.65	⑥⑦
p_4	7	1.76	④⑧⑨⑩

Table 6 Total flight distance

Place of arms	Total flight distance
p_1	11.06
p_2	11.33
p_3	13.82
p_4	14.20

investment, population density and road accessibility, and adopting subjective and objective comprehensive methods to evaluate the site selection of the target area.

The future research will obtain the relationship between the comprehensive assessment model and the real situation, integrating big data into the assessment system. Carrying out network dependency identification and vulnerability assessment to establish the assessment system standards suitable for the national jurisdictions on this basis. It lays a foundation for the future, using drones in response to public emergencies and medical and health emergencies, will performance better than personnel.

Acknowledgements This work is supported by National Key R&D Program of China (No. 2018YFC0807500); National Natural Science Foundation of China 71974067; Key technology R&D Program of Liaoning (2017231005); Social Science Fund Base Project of Beijing, China, No. 19JDGLA001; Fundamental Research Funds for the Central Universities project Number: TW2018009/NTUT-USTB-107-01; National Natural Science Foundation of China, NSFC Project 71602008, 71871016; National natural science foundation of China (No. 71974067).

References

1. C. Zoli, in *Asymmetric Warfare, for the SAGE Encyclopedia of Political Behavior*, ed. by F. Moghaddam, C. Zoli (2016)
2. B. Schuurman, L. Lindekilde, S. Malthaner, F. O'Connor, P. Gill, N. Bouhana, End of the lone wolf: the typology that should not have been. Stud. Conflict Terrorism (2017)
3. C. Zoli, L.J. Steinberg, M. Grabowski, M. Hermann, Terrorist critical infrastructures, organizational capacity and security risk. Saf. Sci. **110**, 121–130 (2018)
4. P. Garcia-Aunon, J.J. Roldán, A. Barrientos, Monitoring traffic in future cities with aerial swarms: developing and optimizing a behavior-based surveillance algorithm. Cogn. Syst. Res. (2018). https://doi.org/10.1016/j.cogsys.2018.10.031
5. Z. Lihua, D. Jianfeng, W. Yu, W. Zhiqiang, An online priority configuration algorithm for the UAV swarm in complex context. Procedia Comput. Sci. **150**, 567–578 (2019)
6. M. Hu, W. Liu, J. Lu, R. Fu, K. Peng, X. Ma, On the joint design of routing and scheduling for Vehicle-Assisted Multi-UAV inspection. Future Gen. Comput. Syst. **94**, 214–223 (2019)
7. A.C. Kapoutsis, S.A. Chatzichristofis, E.B. Kosmatopoulos, DARP: divide areas algorithm for optimal multi-robot coverage path planning. J. Intell. Robot. Syst. **86**(3–4), 663–680 (2017)
8. M. Khanmohamadi, M. Bagheri, N. Khademi, S.F. Ghannadpour, A security vulnerability analysis model for dangerous goods transportation by rail—case study: chlorine transportation in texas-illinois. Saf. Sci. S0925753517313395 (2018)
9. X. Ding, Defense resource optimal allocation problem of sudden terrorism attack incident[J]. Tongji Daxue Xuebao/J. Tongji Univ. **43**(7), 1111–1115 (2015)
10. Z. Corri, L.J. Steinberg, G. Martha, H. Margaret, Terrorist critical infrastructures, organizational capacity and security risk. Saf. Sci. S0925753517305994 (2018)
11. H. Xiaofeng, S. Shifei, Study on the resource allocation in urban defense engineering with intentional threats. Syst. Eng. Procedia **5**, 198–206 (2012)
12. W.A.N.G. Wenming, W.A.N.G. Quanyu, W.A.N.G. Yinghao et al., Research and design of intelligent monitoring and early warning database system for sensitive areas. Netinfo Secur. **19**(12), 1–9 (2019)
13. http://www.chinapeace.gov.cn/chinapeace/c54222/2020-01/17/content_12316856.shtml

14. A.M. Barrett, Mathematical Modeling and Decision Analysis for Terrorism Defense: Assessing Chlorine Truck Attack Consequence and Countermeasure Cost Effectiveness. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania (2009)
15. A.D. Maty, Field guide to tank cars. Bureau of Explosives, Association of American Railroad, Washington, DC (the URL: www.aar.org) (2017)

Selection Strategies on Overseas Logistics Nodes for Chinese Business Logistics Enterprises

Minjie Liang, Zikui Lin, and Kai Zheng

Abstract This paper comprehensively considers whether the construction, operation and management of overseas logistics infrastructure of Chinese enterprises meet the basic conditions from the aspects of the topology and the economic attributes of the Belt and Road trade network. Entropy weight-TOPSIS method is used to comprehensively evaluate the topological importance and economic importance of the national nodes. It comprehensively describes the investment and construction environment of countries along the Belt and Road. Key national nodes in the Belt and Road trade network have been identified. The established investment matrix divides the countries along the Belt and Road into priority, potential and risk investment construction areas, which provides a basis for Chinese business logistics enterprises to "go out" to reasonably choose construction investment priorities and avoid international market risks as much as possible.

Keywords The belt and road · Network topology · Economic attributes · Entropy weight-TOPSIS · International market risk · Investment matrix

1 Introduction

In 2013, President of China proposed the major strategic concepts of developing the Silk Road Economic Belt and 21st-Century Maritime Silk Road for the new century (referred to as the Belt and Road). Since the implementation of the Belt and Road, it has been achieved remarkable results in politics, economy, and culture and other aspects. Promoting the construction of interconnected channels is an essential content

M. Liang · Z. Lin · K. Zheng (✉)
Beijing Jiaotong University, Beijing, China
e-mail: zhengkai@bjtu.edu.cn

M. Liang
e-mail: 19120579@bjtu.edu.cn

Z. Lin
e-mail: zklin@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_62

of the Belt and Road initiative. Countries and regions along the route actively participate in the construction of key nodes and channels in transportation infrastructure, including the China-Laos Railway, the China-Thailand Railway, the Gwadar Port, the Pireaus Port, the China-Russia oil pipeline and the China-Central Asia natural gas pipeline. Business logistics enterprises are encouraged by the Chinese government to "go out" and enhance cooperation of logistics in international regional. International logistics centers have been constructed at major transportation nodes along the Belt and Road to promote the construction of trade channels between countries. However, international operations are not easy for Chinese enterprises. The risks and challenges of politics, finance, market, law, and humanities are encountered in the construction, operation, and management of overseas logistics infrastructure, which may suffer significant losses in case of incorrect response. During the selection of key construction investment projects along the Belt and Road, Chinese enterprises need to consider the satisfaction of investment environment and construction conditions of the nodes besides the key nodes and connections in the Belt and Road trade network.

The innovation and practical values of this paper are reflected in the following aspects. First of all, the topological characteristics of the Belt and Road trade volume weighted network were primarily analyzed in studying the key nodes of "going out" construction investment for Chinese enterprises. The centrality indexes in complex network theory were adopted to identify key national nodes in the network. Meanwhile, the attributes of the nodes were considered whether the logistics infrastructure of primary conditions were met in terms of the construction, operation, and management of the national nodes along the Belt and Road, including logistics performance index, nationality cooperation index, investment environment, information level, and infrastructure development. Then, the work evaluated the topological importance and economic importance of the national nodes through the Entropy weight-TOPSIS. The investment and construction environment of countries along the Belt and Road was described from multiple dimensions of the comprehensive index, aspect index, and regional differences. The research on the investment and construction environment was expanded to a certain extent. Finally, according to the investment matrix established based on the topological importance and economic importance of the national nodes, the countries along the Belt and Road were divided into different construction and investment regions of priority, potential, and risk. The division provides guidance for Chinese business logistics enterprises to rationally select key nodes for construction and investments in the "going out" process, which helps them avoid risks in the international market. Then, the surrounding areas are radiated for the rapid development of the economy and logistics.

2 Literature Review

Complex network theory is widely used to identify key nodes and connections in the network, focusing on the measurement of nodes in the network topology [1, 2]. For example, Zhao [3], Li [4], and Liu [5] adopted central indicators of degree centrality, intermediary centrality, close centrality and eigenvector centrality in the complex network theory. The importance of nodes in various types of road networks is evaluated through the TOPSIS method. The complex unauthorized networks reflect the connection between nodes and the topological characteristics of the network, but it can't describe the strength of the interaction between nodes. Lots of crucial objective information is ignored in the simple topological structure. Thus, edge weights are introduced to characterize the difference in interaction strength between nodes, forming a complex weighted network. Shao [6], Wen [7], Wandelt [8] and Zhang [9] constructed a weighted trade network with the country as the nodes and the import and export trade volume or material flow as the connection. The centrality indicators and HITS algorithm were used to research the key national nodes in the weighted trade network of the world. In a real complex network, nodes attributes, such as economic development level and logistics development capability, have a significant influence on the evolution of the network. For example, Thiyen [10], Feng [11], Liang [12] and Wang [13] established the evaluation indexes of nodes importance from the dimensions of economic development level, infrastructure construction, logistics market status and foreign trade level. The principal component analysis was used to determine the importance of each city node on the international logistics channel. Generally, the identification of key nodes and connections in the network is usually based on the topological characteristics of the nodes, while more attention is paid to the importance of nodes attribute information.

In the study on the investment environment and construction conditions in the Belt and Road, Zhang [14] and Ma [15] used the mean principal component analysis and the factor analysis to compare and evaluate the investment facilitation levels of countries along the Belt and Road. The analysis was conducted from the aspects of economic status, infrastructure, business environment, information technology, financial services, and institutional environment. With the help of the Delphi method, Li [16] evaluated the investment environment of countries along the Belt and Road from six aspects of economic development level, transportation infrastructure construction, informatization level, resource allocation, political environment, and security environment. Xie [17] analyzed the investment environment of Southeast Asian and the Central and Eastern European countries from four dimensions of opening-up level, political system environment, infrastructure construction level, and labor availability. Quan [18], Fang [19] and Zhang [20] adopted entropy weight to evaluate the investment environment of countries along the Belt and Road from the aspects of economic performance, government performance, factor endowments, infrastructure, and institutional environment, and Yin [19] divided the regions into priority, sub-optimal and cautious investment environment levels through systematic cluster analysis. Jiang [21] studied the current status and investment potential of infrastructure along the Belt and Road from the perspectives of transportation infrastructure, communication infrastructure, and power infrastructure. Meanwhile, strategic choices were proposed for cooperation opportunities, external risks, and internal challenges. Therefore, researchers analyze the investment and construction environment of countries along the Belt and Road from different dimensions, where economic development level, political system environment, and infrastructure construction are relatively common evaluation indexes.

3 Trade Network of the Belt and Road

In the network topology $G = \{V, E, M\}$, $V = \{v_i | i \in I\}$ is the set of nodes. n is the total number of nodes in $i = \{1, 2, \ldots, n\}$.

$E = \{e_{ij} = (v_i, v_j) | i, j \in I\}$ is the edge set between node i and node j. $M = (m_{ij})_{n\times n}$ is the adjacency matrix of the network. If there is a connection between node v_i and v_j, then the assigned value of m_{ij} is 1, otherwise, the value is 0. m_{ij} is defined as follows:

$$\begin{aligned} m_{ij} &= 1, (v_i, v_j) \in E \\ m_{ij} &= 0, (v_i, v_j) \notin E \end{aligned} \tag{1}$$

The trade network of the Belt and Road takes the countries along the route as the node set, and trade between them as the edge set. Volume of trade between countries is the edge weight. Table 1 shows the 65 representative countries selected by the node set data from the official website of the Belt and Road. Besides China, there are 2 countries in Northeast Asia, 11 countries in Southeast Asia, 7 countries in South Asia, 20 countries in West Asia and North Africa, 19 countries in Central and Eastern Europe, and 5 countries in Central Asia. The edge set data is from the import and export trade volume in the United Nations Commodity Trade Database in 2018. Figure 1 shows the location of countries along the trade network drawn with Ucinet6.0, according to the latitude and longitude coordinates of the nodes.

Table 1 Geographical distribution of some countries

Region	Countries
Northeast Asia	Mongolia, Russian
Southeast Asia	Singapore, Indonesia, Malaysia, Thailand, Cambodia, Vietnam, Philippines, Myanmar, Laos, Brunei, Timor-Leste
South Asia	India, Pakistan, Bangladesh, Sri Lanka, Maldives, Nepal, Bhutan
West Asia and North Africa	Iran, Iraq, Turkey, Syrian, Jordan, Lebanon, Israel, Palestine, Saudi Arabia, Yemen, Oman, United Arab Emirates, Qatar, Kuwait, Bahrain, Azerbaijan, Georgia, Egypt, Afghanistan, Armenia
Central and Eastern Europe	Poland, Lithuania, Estonia, Latvia, Czech, Slovak, Hungary, Slovenia, Croatia, Bosnia and Herzegovina, Montenegro, Serbia, Albania, Romania, Bulgaria, Macedonia, Ukraine, Belarus, Moldova
Central Asia	Kazakhstan, Uzbekistan, Turkmenistan, Tajikistan, Kyrgyzstan

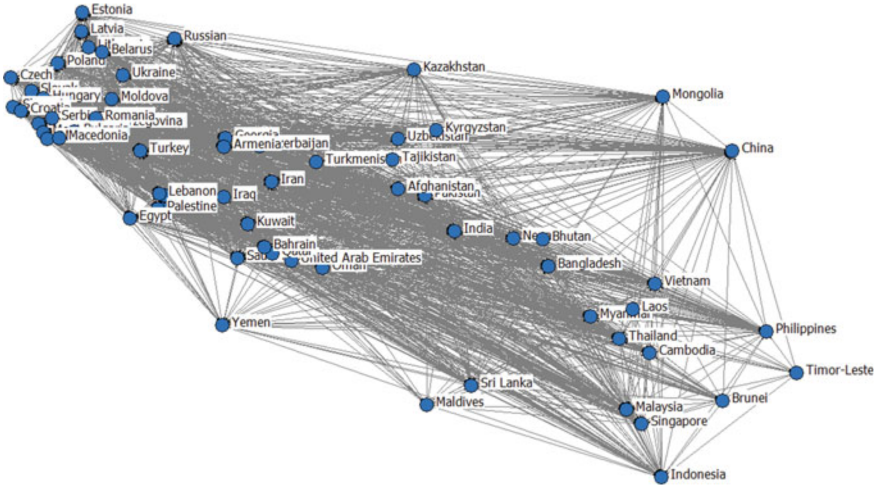


Fig. 1 The Belt and Road trade network

4 Importance Evaluation Index

The selection of key areas for construction and investment along the Belt and Road from a single dimension is one-sided. Therefore, this paper analyzed the topological and economic importance of the national nodes from the topological structure of the network and their economic attributes. The trade links between countries constitute a weighted trade network, and the central indicators are used by the topological importance to measure the key national nodes. However, economic importance reflects the potential of operational logistics infrastructure through the attributes of the national nodes, such as logistics performance index, country cooperation index, national investment environment, information development level, and infrastructure development. Table 2 shows the importance evaluation index system of the nodes

Table 2 Importance evaluation index system of the national nodes

Dimension	Index	Number	Weight
Topological importance	Degree centrality	X1	0.648
	Closeness centrality	X2	0.069
	Betweenness centrality	X3	0.283
Economic importance	Logistics performance	Y1	0.207
	Country cooperation	Y2	0.240
	National investment environment	Y3	0.090
	Information development level	Y4	0.224
	Infrastructure development	Y5	0.239

along the Belt and Road. The index weight is calculated using the entropy weight method, as shown in Eq. (8).

4.1 Topological Importance

The centrality measures of the trade network along the Belt and Road define the importance of nodes through the connections with other nodes. Therefore, this paper selected degree centrality, closeness centrality, and betweenness centrality to assess the topological importance of the national nodes, due to their highly effectiveness on quantifying node importance within a network.

1. *Degree centrality*: Eq. (2) is the degree centrality (α_i) of node i, that is, the ratio of the degree of i (d_i represents the number of nodes directly connected to i) and the maximum number of connections in the network. Greater α_i means the higher importance of i, reflecting the influence of a node on other nodes in the network. In the trade network of the Belt and Road, the degree centrality of the national nodes is directly proportional to the cooperation with other countries along the route.

$$\alpha_i = \frac{d_i}{n-1} \tag{2}$$

2. *Closeness centrality*: Eq. (3) shows the closeness centrality (β_i) of i, which is the reciprocal of the average distance along the shortest path from i to other nodes. d_{ij} represents the shortest path between node i and j. Greater β_i means the higher importance of i, namely i has a more significant impact on other nodes. In the trade network of the Belt and Road, the higher centrality of a national node implies that the country in a central position is more comfortable to conduct international trade with other countries.

$$\beta_i = \frac{n-1}{\sum_j^n d_{ij}} \tag{3}$$

3. *Betweenness centrality*: Eq. (4) shows the betweenness centrality (γ_i) of i, which is the ratio of the number of shortest paths of i between j and k. g_{jk} represents the number of shortest paths between j and k, while g_{jk} represents that between j and k through i. The larger the γ_i, the higher the control of i to other nodes. In the trade network of the Belt and Road, higher betweenness centrality of the 13 second-level indicators, and 28 third-level indicators. Sorting out the advantages and disadvantages of each national nodes means that it has a more reliable control in the principal channel of the trade network.

$$\gamma_i = \frac{2}{n^2 - 3n + 2}\sum_j^n \sum_k^n \frac{g_{jk}^i}{g_{jk}}, j \neq k \neq i, j < k \tag{4}$$

4.2 Economic Importance

In the literature, economic development, market environment, infrastructure, logistics development, and informatization level are high-frequency indexes to evaluate the importance of nodes and the investment and construction environment. However, considering the difficulty of quantifying the indexes and the availability of data, this paper selected logistics performance index, national cooperation index, national investment environment, information development level, and infrastructure development to evaluate the economic importance of the national nodes in the trade network.

1. *Logistics performance index (LPI)*: The LPI comes from the Connecting to Competitiveness-Trade Logistics in the Global Economy, which is compiled by the World Bank's International Trade Department every two years. The comprehensive logistics capabilities of countries are reflected through the customs, infrastructure, international shipments, logistics quality and competence, tracking and tracing, timeliness.
2. *Country cooperation index*: The country cooperation index comes from the Big Data Report of the Belt and Road, putting forward five key points of cooperation around the Vision and Actions on Jointly Building Silk Road Economic Belt and 21st-Century Maritime Silk Road. The evaluation index system has been constructed from the policy connectivity, infrastructure connectivity, trade connectivity, financial connectivity, and people to people connectivity, including 5 first-level indicators, 12 second-level indicators, and 34 third-level indicators, which reflects the effectiveness of cooperation between China and the countries along the route.
3. *National investment environment*: The national investment environment comes from the Big Data Report of the Belt and Road, reflecting the investment environment of countries along the route. It has constructed an evaluation index system from the political environment, economic environment, business environment, natural environment, and relations with China, including 5 first-level indicators country's investment environment can help companies explore opportunities and avoid risks in investment cooperation.
4. *Information development level*: The level of information development comes from the Big Data Report of the Belt and Road, which builds an evaluation index system including 3 first-level and 11 second-level indicators from ICT foundation, applications, and industry. The penetration rate of information and communication products and the external supply capacity of products and services reflect the informatization development.

5. *Infrastructure development*: The infrastructure development index comes from the National Infrastructure Development Index Report of the Belt and Road. Referring to the classifications of the World Bank and BMI, infrastructure is divided into transportation, energy, utility services, and buildings. Transportation facilities include roads, railways, airports, ports. Energy refers to oil and gas and electricity, and the utility services are provided through water conservancy projects and communication networks. Besides, buildings are used for civil, industrial, and commercial purposes, reflecting the country's level of infrastructure development.

5 Comprehensive Evaluation Method

The topological and economic importance indexes of a national node can indicate its importance in the trade network of the Belt and Road. However, it is limited to specific aspects, and more comprehensive features are obtained with the entropy weight-TOPSIS. This method compares the limited objects with idealized objects to evaluate the existing objects. The necessary steps are as follows.

5.1 Eigenfactor Matrix of Nodes

$F = \{f_1, f_2, f_3, f_4\}$ represents four evaluation indexes. $v_i(f_i), i = 1, 2, \ldots, n, j = 1, 2, 3, 4$ in the national node set $V = \{v_1, v_2, v_3, v_4\}$ represents the jth importance index of node i (as the characteristic factor). Equation (5) shows the eigenfactor matrix P of the national nodes.

$$P = \begin{bmatrix} v_1(f_1) & v_1(f_2) & v_1(f_3) & v_1(f_4) \\ v_2(f_1) & v_2(f_2) & v_2(f_3) & v_2(f_4) \\ \vdots & \vdots & \vdots & \vdots \\ v_n(f_1) & v_n(f_2) & v_n(f_3) & v_n(f_4) \end{bmatrix} \tag{5}$$

5.2 Matrix Standardization

Considering the value scales of the four indicators are different, the eigenfactor matrix of the national nodes need to be standardized. In Eq. (6), t_{ij} is the jth element of the ith row in the standardized matrix.

$$t_{ij} = \frac{v_i(f_j)}{\sqrt{\sum_{i=1}^{n} v_i(f_j)^2}}, j = 1, 2, 3, 4 \tag{6}$$

5.3 Weighting Eigenfactor Matrix for the Nodes

The four indexes represent different characteristics of the national nodes in the trade network of the Belt and Road, which can be weighted by their importance for balance. If $w_j(j = 1, 2, 3, 4)$ is the jth weighted coefficient of the eigenfactor, Eq. (7) shows the normalized eigenfactor matrix R of the weighted nodes.

$$R = (r_{ij}) = (w_j t_{ij}) \begin{bmatrix} w_1 t_{11} & w_2 t_{12} & w_3 t_{13} & w_4 t_{14} \\ w_1 t_{21} & w_2 t_{22} & w_3 t_{23} & w_4 t_{24} \\ \vdots & \vdots & \vdots & \vdots \\ w_1 t_{n1} & w_2 t_{n2} & w_3 t_{n3} & w_4 t_{n4} \end{bmatrix} \tag{7}$$

This paper obtained the weight of each index with the entropy-weight method to exclude the subjectivity of index weight. The necessary steps are as follows: Assuming that there are m plans to be evaluated, n evaluation indexes form the original index data matrix $X{=}(x_{ij})_{m\times n}$. A more significant distance of the value x_{ij} of the index x_j indicates a higher effect on the comprehensive evaluation. In information theory, information entropy represents the order degree of the system, while the two are proportional. The index weight can be calculated with information entropy according to the difference of each index value. Table 2 shows the index weight of topological and economic importance.

$$e_i = -k \sum_{i=1}^{m} p_{ij} \ln(p_{ij}) \tag{8}$$

5.4 Calculation of the Ideal Values for Index Importance

Equation (9) shows the positive and negative ideal values A^+ and A^-, which are calculated by the maximum and minimum values of each column of the weighted eigenfactor matrix R.

$$\begin{aligned} A^+ &= \{\max(r_{i1}), \ldots, \max(r_{i4})\} = \left\{r_1^+, r_2^+, r_3^+, r_4^+\right\}, i \in \{1, 2, \ldots, n\} \\ A^- &= \{\min(r_{i1}), \ldots, \min(r_{i4})\} = \left\{r_1^-, r_2^-, r_3^-, r_4^-\right\}, i \in \{1, 2, \ldots, n\} \end{aligned} \tag{9}$$

5.5 Distance Calculation

Equation (10) shows the distance between the national node i and the positive and negative ideal values (A^+ and A^-), which is calculated by the relative-entropy distance model.

$$
\begin{aligned}
S_i^+ &= \left\{ \sum_{j=1}^{4} \left[r_j^+ \log \frac{r_j^+}{r_{ij}} + (1 - r_j^+) \log \frac{1 - r_j^+}{1 - r_{ij}} \right] \right\}^{\frac{1}{2}} \\
S_i^- &= \left\{ \sum_{j=1}^{4} \left[r_j^- \log \frac{r_j^-}{r_{ij}} + (1 - r_j^-) \log \frac{1 - r_j^-}{1 - r_{ij}} \right] \right\}^{\frac{1}{2}}
\end{aligned}
\tag{10}
$$

5.6 Closeness Calculation

Equation (11) shows that C_i can evaluate the importance of the national nodes. Greater C_i means node i is farther from the negative ideal value and closer to the positive ideal value.

$$
C_i^* = \frac{S_i^-}{S_i^+ + S_i^-}, i = 1, 2, \ldots, n \tag{11}
$$

6 Topological and Economic Importance Analysis

6.1 Node Importance Analysis

The topological importance and economic importance scores and rankings of national nodes are shown in Table 3. In terms of regions, the top-ranked countries in Southeast Asia account for the majority. Singapore is particularly prominent, with its economy, politics, infrastructure, logistics, and information technology at the forefront of the world. Singapore, Malaysia, and Indonesia are the most critical energy-transportation channels because of their strategic position on the brink of the Strait of Malacca.

The top rankings in South Asia are India and Pakistan, and India with the highest topological importance except Russia, but the economic importance is only ranked 19th. As one of the fastest-growing countries in the world, India is at the center of the trade network of the Belt and Road for the rapid economic growth and infrastructure

Table 3 Topological and economic importance scores and rankings of the national nodes

Topological importance						Economic importance					
Country	Score	Ranking	Country	Score	Ranking	Country	Score	Ranking	Country	Score	Ranking
Russian	0.314	1	Bangladesh	0.107	16	Singapore	0.839	1	Romania	0.589	16
India	0.311	2	Egypt	0.102	17	Russian	0.804	2	Philippines	0.586	17
Malaysia	0.303	3	Iran	0.098	18	Malaysia	0.722	3	Pakistan	0.584	18
Singapore	0.301	4	Hungary	0.097	19	Poland	0.719	4	India	0.584	19
United Arab Emirates	0.259	5	Sri Lanka	0.095	20	Thailand	0.690	5	Belarus	0.544	20
Vietnam	0.255	6	Iraq	0.089	21	United Arab Emirates	0.681	6	Cambodia	0.535	21
Thailand	0.250	7	Bulgaria	0.083	22	Indonesia	0.671	7	Bahrain	0.523	22
Indonesia	0.237	8	Ukraine	0.080	23	Israel	0.668	8	Mongolia	0.523	23
Saudi Arabia	0.210	9	Romania	0.077	24	Kazakhstan	0.666	9	Egypt	0.521	24
Turkey	0.185	10	Kuwait	0.071	25	Vietnam	0.660	10	Estonia	0.510	25
Poland	0.167	11	Jordan	0.066	26	Turkey	0.648	11	Serbia	0.501	26
Czech	0.138	12	Lebanon	0.064	27	Czech	0.646	12	Kuwait	0.495	27
Philippines	0.126	13	Oman	0.059	28	Hungary	0.610	13	Bulgaria	0.487	28
Pakistan	0.114	14	Qatar	0.056	29	Saudi Arabia	0.604	14	Kyrgyzstan	0.483	29
Slovak	0.112	15	Kazakhstan	0.056	30	Qatar	0.600	15	Laos	0.481	30
Bangladesh	0.107	16	Moldova	0.020	48	Romania	0.589	16	Azerbaijan	0.376	48
Egypt	0.102	17	Macedonia	0.017	49	Philippines	0.586	17	Lebanon	0.376	49
Iran	0.098	18	Armenia	0.016	50	Pakistan	0.584	18	Tajikistan	0.370	50
Hungary	0.097	19	Syrian	0.014	51	India	0.584	19	Macedonia	0.351	51

(continued)

Table 3 (continued)

Topological importance						Economic importance					
Country	Score	Ranking	Country	Score	Ranking	Country	Score	Ranking	Country	Score	Ranking
Sri Lanka	0.095	20	Kyrgyzstan	0.014	52	Belarus	0.544	20	Armenia	0.343	52
Iraq	0.089	21	Afghanistan	0.013	53	Cambodia	0.535	21	Moldova	0.326	53
Bulgaria	0.083	22	Mongolia	0.013	54	Bahrain	0.523	22	Bosnia and Herzegovina	0.323	54
Ukraine	0.080	23	Turkmenistan	0.013	55	Mongolia	0.523	23	Montenegro	0.316	55
Romania	0.077	24	Brunei	0.013	56	Egypt	0.521	24	Albania	0.298	56
Kuwait	0.071	25	Laos	0.010	57	Estonia	0.510	25	Turkmenistan	0.289	57
Jordan	0.066	26	Yemen	0.009	58	Serbia	0.501	26	Iraq	0.265	58
Lebanon	0.064	27	Tajikistan	0.008	59	Kuwait	0.495	27	Afghanistan	0.234	59
Oman	0.059	28	Montenegro	0.004	60	Bulgaria	0.487	28	Timor-Leste	0.233	60
Qatar	0.056	29	Palestine	0.002	61	Kyrgyzstan	0.483	29	Palestine	0.191	61
Kazakhstan	0.056	30	Maldives	0.002	62	Laos	0.481	30	Bhutan	0.189	62
Israel	0.055	31	Bhutan	0.001	63	Iran	0.477	31	Yemen	0.126	63
Belarus	0.053	32	Timor-Leste	0.000	64	Slovenia	0.476	32	Syrian	0.115	64

construction. Pakistan ranks in the top 20 in terms of topological and economic importance, whose close cooperation with China is known as the "blood brother".

The five countries in Central Asia are not ranked in the top 20 in terms of topology and economic importance simultaneously. Kazakhstan, the world's largest landlocked country and the connection channel between Europe and Asia, is rich in mineral resources and ranks ninth in economic importance. It is also a critical transit-point for Chinese business logistics enterprises to radiate westward.

Saudi Arabia, the UAE, and Turkey rank high in West Asia and North Africa. Saudi Arabia and the UAE are located in the Arabian Peninsula near the Persian Gulf and the Strait of Hormuz, rich in oil and natural gas resources. As the only maritime channel for oil in the Gulf to all parts of the world, they are the hub of culture, economy, and trade between East and West countries. Being the crossroads of Europe and Asia, the Turkish Strait is the only channel connecting the Black Sea and the Mediterranean Sea. Moreover, its geographical location and strategic geopolitical significance are essential. Demand for infrastructure construction in the Middle East is eager. Although the Belt and Road Initiative has played an active role, there are potential obstacles, including religious sect disputes, terrorist activities, political turmoil, sanctions by Western countries, and national political intervention.

The top European rankings are Russia, Poland, and the Czech Republic, of which Russia ranks first or second in terms of topological and economic importance. Russia is a crucial link connecting Asia and Europe, actively responding to the Belt and Road initiative and bordering many countries along the route. Poland and the Czech Republic are the top-ranking countries in terms of topological and economic importance in Central and Eastern Europe. Located at the intersection of Europe's "Amber Road" and "Silk Road", the geographical location of the two countries can radiate entire Europe. The participation of more European countries in the Belt and Road has broadened the trading partners and deepened the cooperation and sharing among Asian and European countries.

6.2 Regional Importance Analysis

The topological importance index scores and rankings of regions are shown in Table 4. In terms of regional topological importance, Northeast Asia ranks first among all the regions with an average score of 0.164. The reason is that only Mongolia and Russia are included in Northeast Asia, where Russia has the highest topological importance. In terms of sub-items, the degree centrality, closeness centrality, and betweenness centrality scores of Northeast Asia are relatively higher. It indicates that Northeast Asia is closely connected with other regions in the Belt and Road trade network, more likely to building links with other regions. Primarily, Russia crosses the Eurasian continent to border many countries along the Belt and Road, with substantial control over other regions.

Table 4 Regional topological importance scores and rankings

Region	X1	X2	X3	Average score
Northeast Asia	0.120	0.125	0.105	0.164
Southeast Asia	0.098	0.122	0.128	0.143
South Asia	0.045	0.121	0.122	0.094
West Asia and North Africa	0.041	0.120	0.085	0.073
Central Asia	0.015	0.113	0.022	0.023
Central and Eastern Europe	0.030	0.128	0.080	0.059

Southeast Asia ranks second with an average score of 0.143, where Malaysia, Singapore, Vietnam, Thailand, and Indonesia all rank in the top ten. The closeness centrality and betweenness centrality are higher in Southeast Asia. The degree centrality is lower than that in Northeast Asia, which is the main reason for the topological importance second to Northeast Asia. There are thousands of years of shared history and trade exchanges between Southeast Asia and China, with the advantages of a similar geographical location. The Strait of Malacca, as the most critical energy-transportation channel for China, is the "marine lifeline" and the thoroughfare connecting the Pacific and the Indian Ocean. South Asia ranks third with an average score of 0.094, similar to Southeast Asia. South Asia has high closeness centrality and betweenness centrality and low degree centrality, where India is only second to Russia in topological importance.

The topological importance scores are similar among West Asia and North Africa, as well as Central and Eastern Europe. Different form Northeast Asia and Southeast Asia, the betweenness centrality and degree centrality are low in these regions. West Asia, North Africa, and Central and Eastern Europe are far away from China, so they have fewer connections with other regions in the Belt and Road network, with weaker control over other regions. Central Asia has the lowest ranking of topological importance, with a score of 0.023. The degree centrality, closeness centrality, and betweenness centrality are the lowest due to the backward level of economic development and the single structure in Central Asia. As a result, the topological importance of Central Asia is relatively lower in the network.

The economic importance index scores and rankings of regions are shown in Table 5. In terms of regional economic importance, Northeast Asia tops the list with an average score of 0.663. The cause is that Russia ranks second in economic importance, only after Singapore. In terms of sub-items, the nationality cooperation index in Northeast Asia is significantly higher than that in other regions, indicating that it is the most active in the participation and integration of the Belt and Road initiative. The China-Mongolia-Russia Economic Corridor connects the Chinese Silk Road Economic Belt with the Russian Trans-Eurasian Railway and the Mongolian Grassland Road Initiative, with close contact, cooperation, and sharing. Southeast Asia ranks second with an average score of 0.573, and the countries occupy five seats in the top ten in the ranking of the national nodes economic importance. Singapore, Malaysia, and Thailand rank the first, third, and fifth, respectively. However, the level

Table 5 Regional economic importance scores and rankings

Region	Y1	Y2	Y3	Y4	Y5	Average score
Northeast Asia	0.113	0.196	0.137	0.138	0.131	0.663
Southeast Asia	0.127	0.156	0.132	0.107	0.144	0.573
South Asia	0.114	0.127	0.113	0.082	0.124	0.434
West Asia and North Africa	0.121	0.103	0.117	0.112	0.117	0.425
Central Asia	0.111	0.139	0.118	0.077	0.116	0.439
Central and Eastern Europe	0.132	0.089	0.129	0.152	0.115	0.474

of informatization in Southeast Asia is only higher than that in South and Central Asia, which restricts the economic importance of Southeast Asia to a certain extent.

The average scores of economic importance are similar among Central and Eastern Europe, Central Asia, South Asia, as well as West Asia and North Africa, with the values all below 0.5. Except for South Asia, Poland in Central and Eastern Europe ranks fourth, Kazakhstan in Central Asia ranks ninth, and the UAE in West Asia and North Africa ranks sixth. They have forefront rankings of national economic importance. Also, Central and Eastern Europe have the highest informatization level score and the lowest country cooperation index. There are many developed countries in Central and Eastern Europe, with relatively higher economic development level. However, they are far from China geographically and susceptible to other international factors, with low acceptance on the Belt and Road initiative. Central Asia is the region with the highest foreign cooperation index in Northeast Asia, but with the lowest level of informatization. Due to the short distance from China, the countries actively respond to the Belt and Road initiative. However, the East Asian countries are inland with a relatively backward economy and low penetration rate of informatization. Similar to Southeast Asia, the low level of informatization in South Asia restricts economic importance. There is no significant difference between indicators of West Asia and North Africa, while they have the lowest economic importance scores in all regions. Wherein the nationality cooperation index is only higher than that of Central and Eastern Europe, which requires further improvement.

7 Strategies of Investment and Construction

There are lots of countries along the Belt and Road, with various development levels. During the construction and operation of overseas logistics infrastructure, the enterprises need to consider that the national nodes are at the center of the trade network of the Belt and Road to radiate the neighboring countries. Meanwhile, the national nodes need to have a stable business and investment environment as well as strong logistics development capabilities. The corresponding investment strategies should

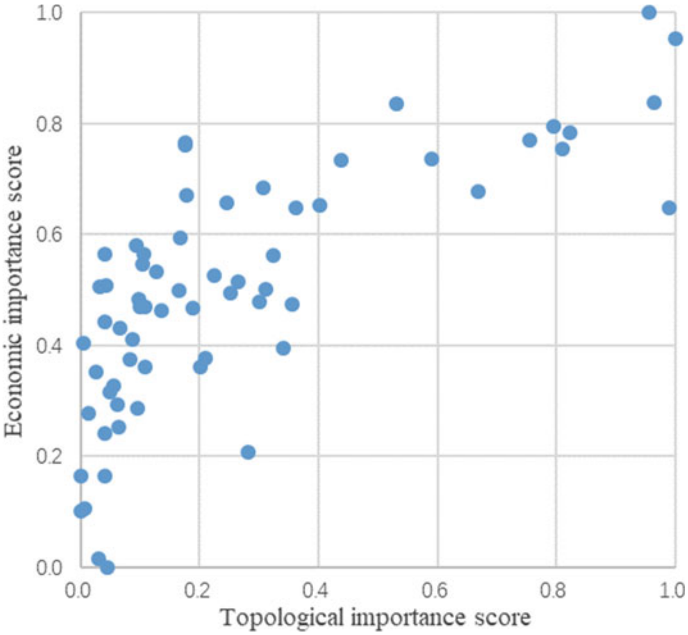


Fig. 2 Topological importance and economic importance standardized scores of the national nodes

be formulated by "going out" Chinese business logistics enterprises in terms of countries with different topological and economic importance rankings in the Belt and Road trade network.

Figure 2 shows the standardized score distribution of topological importance and economic importance of the national nodes. The countries with higher topological importance scores have higher economic importance scores. However, those with higher economic importance scores may not have higher topological importance. The reason is that the national nodes with high topological importance are at the center of the trade network of the Belt and Road, with more robust control over other nodes. They more easily conduct international trade with other countries to promote the development on economic strength, market environment, logistics capabilities, and infrastructure of the countries to a certain extent. According to the definition of "segmentation point" proposed by Kim [22], if the score of the indicator drops significantly at a specific value, then the value before such value is the segmentation point. The element with the score higher than the segmentation point is the core element. The national nodes in the Belt and Road trade network are divided into three regions (see Table 6). The corresponding investment strategies are formulated according to the characteristics of each regional logistics market.

According to the Boston Matrix, all the products of the enterprises are divided into "stars," "question marks", "cash cow", and "dogs" based on the sales growth rate and relative market share. Corresponding decisions are made according to the

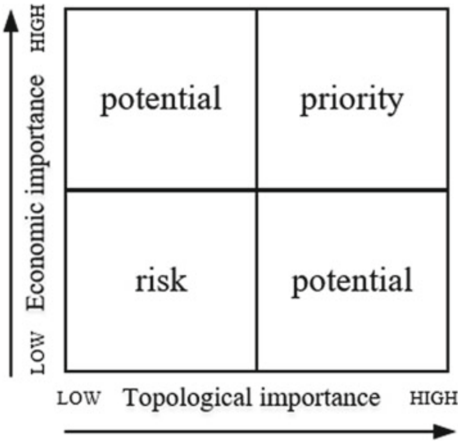
Table 6 Regional distribution of the national nodes

Region	Countries
High topological importance High economic importance	Russian, India, Malaysia, Singapore, United Arab Emirates, Vietnam, Thailand, Indonesia, Saudi Arabia, Turkey, Poland, Czech, Philippines, Pakistan
Low topological importance High economic importance	Israel, Kazakhstan, Hungary, Qatar, Romania, Belarus, Cambodia, Bahrain, Mongolia, Egypt, Estonia, Serbia, Kuwait, Bulgaria, Kyrgyzstan, Laos, Iran
Low topological importance Low economic importance	Slovenia, Ukraine, Lithuania, Sri Lanka, Slovak, Croatia, Latvia, Oman, Myanmar, Brunei, Georgia, Nepal, Maldives, Bangladesh, Jordan, Uzbekistan, Azerbaijan, Lebanon, Tajikistan, Macedonia, Armenia, Moldova, Bosnia and Herzegovina, Montenegro, Albania, Turkmenistan, Iraq, Afghanistan, Timor-Leste, Palestine, Bhutan, Yemen, Syrian

characteristics of different types of products. The products without development prospects are eliminated to achieve a virtuous cycle of product and resource allocation structure. From the basic principles of the Boston Matrix, the investment matrix for Chinese business logistics enterprises was established based on the topological importance and economic importance (see Fig. 3).

The countries with top rankings in terms of topological importance and economic importance are in the center of the trade network of the Belt and Road. They have close trade relations and large trade volume with the countries along the route. The investment environment and business environment are stable in these countries, with strong comprehensive logistics capabilities, a high level of informatization, rapid infrastructure development, and small political and legal risks. They are the priorities for the construction and investment of Chinese business logistics enterprises. However, the logistics market space of some countries has become saturated with

Fig. 3 Investment matrix



higher market entry barriers, thus leading to the risks of high labor costs, land costs, and environmental costs.

The countries with high economic importance and low topological importance have less trade relations and smaller trade volume with the countries along the route. Due to a favorable investment environment and business environment, strong logistics development capability, and rapid infrastructure development, the countries have plenty of development space and strong growth potential in the logistics market and infrastructure. Primarily, there is considerable logistics demand in infrastructure construction fields of energy, transportation, telecommunications, and ports. They are potential regions for construction and investment of Chinese business logistics enterprises.

The countries with low topological importance and economic importance are inferior in the trade network of the Belt and Road. They are susceptible to the influence and control by the nodes in other countries, with less cooperation with China, reduced level of logistics development, slow infrastructure construction as well as unstable investment environment and business environment. Some countries are in geopolitically disordered regions, lagging in the development of economy, politics, infrastructure, logistics, and informatization. The construction and investment in such countries may encounter higher risks.

Acknowledgements Supported by "the Fundamental Research Funds for the Central Universities" (2017JBZ005).

References

1. J. Liu, Z. Ren, Q. Guo, B. Wang, Research progress of nodes importance ranking in complex networks. J. Phys. **62**(17), 9–18 (2013)
2. L. Xiao, L. Lv, Review of ranking methods of important nodes on the network. Sci. Bull. **59**(13), 1175–1197 (2014)
3. L. Zhao, Y. Zhao, Q. Hu, H. Li, J. Stoeter, Evaluation of consolidation center cargo capacity and locations for China railway express. Transp. Res. Part E: Logist. Transp. Rev. **117**, 58–81 (2018)
4. D. Li, L. Zhao, C. Wang, W. Sun, J. Xue, Selection of China's imported grain distribution centers in the context of the Belt and Road initiative. Transp. Res. Part E: Logist. Transp. Rev. **120**, 16–34 (2018)
5. H. Liu, Y. Ru, B. Sun, Research on passenger transport capacity of national provincial capital city railway network. Econ. Geogr. **36**(08), 16–22 (2016)
6. Z. Shao, Z. Ma, J. Sheu, H. Gao, Evaluation of large-scale transnational high-speed railway construction priority in the belt and road region. Transp. Res. Part E: Logist. Transp. Rev. **117**, 40–57 (2018)
7. X. Wen, H. Ma, T. Choi, J. Sheu, Impacts of the belt and road initiative on the China-Europe trading route selections. Transp. Res. Part E: Logist. Transp. Rev. **122**, 581–604 (2019)
8. S. Wandelt, X. Sun, Evolution of the international air transportation country network from 2002 to 2013. Transp. Res. Part E: Logist. Transp. Rev. **82**, 55–78 (2015)
9. C. Zhang, J. Fu, Z. Pu, A study of the petroleum trade network of countries along "The Belt and Road Initiative". Transp. Res. Part E: Logist. Transp. Rev. **222**, 593–605 (2019)

10. N. Thiyen, J. Zhang, Y. Yan, Research on ASEAN regional international logistics nodes[J]. Compr. Transp. **40**(07), 97–102 (2018)
11. R. Feng, *International Logistics Channel Construction Method and Application Research* (Southwest Jiaotong University, 2017)
12. C. Liang, X. Liu, Y. Gong et al., Construction of a mixed hub-and-spoke logistics network in Beijing-Tianjin-Hebei multi-hub. China Circ. Econ. **33**(06), 118–126 (2019)
13. R. Wang, Construction of regional logistics network based on axial-spoke theory——empirical test based on Shandong provincial data. Bus. Econ. Res. **06**, 91–94 (2018)
14. Y. Zhang, "Belt and Road" investment facilitation and China's foreign direct investment choices-an empirical study based on cross-country panel data and investment gravity model. Int. Trade Issues **09**, 165–176 (2016)
15. W. Ma, M. Qiao, Measurement and Evaluation of the "Belt and Road" national investment facilitation level. J. Hebei Univ. (Philos. Soc. Sci. Ed.) **41**(05), 85–94 (2016)
16. Y. Li, J. Zheng, X. Jin et al., Comprehensive assessment and countermeasures for the investment environment of the "Belt and Road". Bull. Chin. Acad. Sci. **31**(06), 671–677 (2016)
17. G. Xie, Y. Xu, F. Yang, Comparative research on the investment environment of southeast Asian and Central and Eastern European Countries under the Background of "One Belt and One Road". World Econ. Res. **2018**(11), 89–98+137 (2018)
18. H. Quan, P. Zhang, Analysis of the "Belt and Road" construction and corporate investment environment in Asia. J. Shanghai Univ. Financ. Econ. **19**(01), 88–102 (2017)
19. Y. Fang, J. Chen, H. Dai, Comprehensive evaluation of the investment environment of the Gulf countries under the background of the "Belt and Road". World Geogr. Res. **27**(02), 36–44+94 (2018)
20. C. Zhang, H. Man, Comprehensive evaluation and comparison of the investment environment of countries along the "Belt and Road"—an empirical study based on different types of economies. Financ. Econ. **02**, 48–54 (2018)
21. W. Jiang, Infrastructure investment and construction along the "Belt and Road" and China's strategic choices. Int. Trade **12**, 44–52 (2017)
22. Y. Kim, T.Y. Choi, T. Yan, Structural investigation of supply networks: a social network analysis approach. J. Oper. Manage. **29**(3), 194–211 (2011)

Evaluation Model Construction of Brand Media Influence in Short Video in the Publishing Media Industry

Shaozhen Hong and Liang Wang

Abstract Facing the impact of the current digital wave, the Publishing media industry is in urgent need of exploring a new way for brand communication. This paper provides a model for publishing media industry to quantify its own brand communication influence. Based on the perspective of brand communication characteristic, this paper combines with the feature acquisition principle of mobile terminal to redefine brand communication elements. Four research questions are proposed, and quantitative indicators of brand evaluation with calculation methods are constructed. Compared with NEWRANK index, this paper illustrates the feasibility of the evaluation model of the brand media influence of publishing media in short videos. At last, three suggestions are put forward to help publishing media enterprises to explore a new way to quantify brand media influence.

Keywords Publishing and media industry · Short video · Brand media influence · Evaluation model

1 Introduction

From "digital publishing" to "transformation and upgrading" according with "knowledge services" to "integrated development", the critical period of climbing the slope and overcoming obstacles will take place in 2020. The application of 5G technology will profoundly influence and change the pattern of content organization, production mode and content dissemination [1]. The traditional "editing, publishing, distribution" has not covered the connotation of publishing activities. The publishing media industry has experienced from lead and fire to light and electricity, and then it gets to the combination of digital and network. Facing the wave of "digital era" shock,

S. Hong · L. Wang (✉)
School of Economics and Management, Beijing Institute of Graphic Communication, Beijing, China
e-mail: wangliang@bigc.edu.cn

S. Hong
e-mail: 15717260757@163.com

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_63

the publishing media industry needs to use the granular marketing of short video to build the brand media influence so as to create market value and occupy the market efficiently. It seems evitable for publishing media industry to establish connection with big data, cloud computing, AI and other technologies. In order to expand brand influence, this mode of transmission will significantly expand the application value and influence of digital publishing. Brands do serve as important signals of quality, meaning, and value [2].

Since 2010, the international digital publishing environment has undergone drastic changes. The Publishing media industry in China completed the transformation into an enterprise that year, and the annual number of printed books decreased significantly in the United States, Britain, Russia, South Korea and other countries. Digital publishing has already become a new interest growth point in the publishing industry [3]. In recent years, statistics on the digital publishing industry by the China academy of press and publication since 2006 have shown that the total revenue of the digital publishing industry has increased by eight times from 21.3 billion yuan in 2006 to 193.55 billion yuan in 2012, with an average annual growth rate of over 40%. According to *The 2019 Annual Report on Content Entrepreneurship*, the number of users of short video apps has reached 594 million, accounting for 74.19% of the total number of Internet users. On the early morning of June 12th, 2019, Queen of the Internet—*Mary Meeker* released *The Internet Trends Report 2019*, which showed that from April 2017 to April 2019, the average daily use time of short video apps in China increased from less than 100 million hours to 600 million hours. Take TikTok as an example. In 2017, TikTok became a hot platform with over 150 million daily active users and over 300 million monthly active users. Most of the users are between 25 and 30 years old, accounting for 29.13%. Users of the whole platform are relatively young, and the ratio of male to female is close to 1:1. According to the survey report, about 20 percent of TikTok users play TikTok for more than an hour every day [4]. People use TikTok not only for entertainment, but also for keeping up with “fashion” and the informational and practical needs [5]. Such strong user engagement has made short video one of the most influential promotional media in recent years.

This paper embarks from the short video media. It explores the brand publication medium enterprise in a short video transmission influencing evaluation model, providing a feasible evaluation system for digital publishing media industries. It also speeds up the construction of integrated and provides one full media publishing communication pattern, which promotes the publishing industry up to a new level. This paper illustrates the effect of the depth and breadth of communication on the brand’s communication influence with the wave lobe chart and the influence radiation chart, which are innovative to some extent. In addition, this paper conducts a quantitative analysis of brand communication influence, which calculates the top ten ranking lists of brand influence of Chinese publishing and media enterprises with evaluation indexes. Lastly, it verifies the feasibility of the quantitative calculation model of brand communication influence proposed in this paper with the *NEWRANK* index. The specific research idea is shown as Fig. 1.

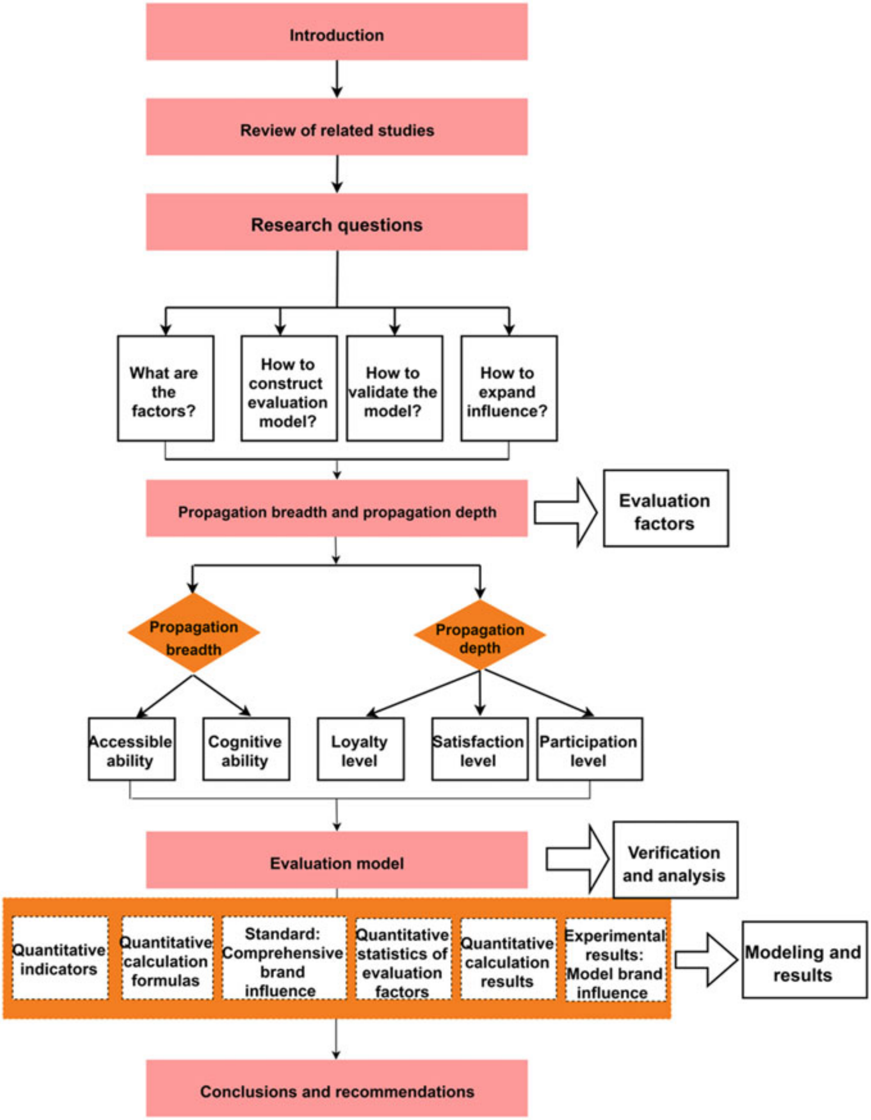

Introduction
Review of related studies
Research questions
What are the factors?
How to construct evaluation model?
How to validate the model?
How to expand influence?
Propagation breadth and propagation depth
Evaluation factors
Propagation breadth
Propagation depth
Accessible ability
Cognitive ability
Loyalty level
Satisfaction level
Participation level
Evaluation model
Verification and analysis
Quantitative indicators
Quantitative calculation formulas
Standard: Comprehensive brand influence
Quantitative statistics of evaluation factors
Quantitative calculation results
Experimental results: Model brand influence
Modeling and results
Conclusions and recommendations

Fig. 1 Flow chart of the research idea

2 Review of Related Studies

Since the beginning of the era of diversified perspectives in brand research in the late 1990s, Chinese publishing and media enterprises have undergone significant changes in market competition gradually and the content industry was showing a trend of cross-media convergence. A brand is the guarantee of a company's continuous profits, and it is also the commercial asset that stands the most difficulty for an enterprise to be copied. From products and services to an enterprise, it needs the support of a strong brand [4]. Throughout the discussion on "brand" in the academic circle, scholars believe that the research process of brand theory has gone through the following stages: in the first stage, brand concept began to sprout, and individual brand research started before 1915; The second stage, 1915–1945, brand management and brand research started; In the third stage, from the 1950s to the 1980s, the systematic study of brand theory was advanced. Burleigh b. Gardner and Burleigh B. Gaxdner and Sjdney. Levy published the brand research paper *Products and Brands*, which was the earliest systematic study of brand theory. Rosser Reeves first linked advertising to brands in the early 1950s with his Unique Selling Proposition (USP). David Oglny, an advertising master who was active in the 1950s and 1960s, put forward the famous Brand Image theory. The fourth stage, from the late 1980s to 1990s, is an integrated brand management theory study time. The main views are research of one of the brand authoritative scholar in the field of study in the United States by David Ike (Avid. Aaker) put forward "brand trilogy"—"managing brand equity" (1991), "establish a strong brand" (1996), and "brand leadership" (2000). In addition, professor Kevin lane Keller, a famous Brand management strategy expert, proposed the Brand Equity model based on Customer value (customer-based Brand Equity) in his monumental book *Strategic Brand Management* (1995). The fifth stage, from the end of the 1990s to the present, is the era of diversified brand research perspectives. During this period, David was introduced to the concept of population ecology study of the theory of brand for the first time. Based on single enterprise brand "brand colony" concept, he put forward the system and built the relationship between the ecological brand. At this point, the publishing industry, truly built up the relationship between the enterprise brand with the brand market, brand, customers/stakeholders, resources, environment and the system.

Since 2003, *the General Administration of Press and Publication of China* proposed carrying out the "brand project" to build a batch of famous press, famous newspapers, famous journals and high-quality publications. Based on the definition of media brand by scholars the author concludes as follows: the brand of publishing media enterprise refers to the sum of the reputation, credibility and communication power provided by publishing media enterprise in the eyes of audiences. The specific performance is: Including the book commodity and the service function element, the press and the book commodity image element and the reader psychology element three-dimensional synthesis.

At present, researches on media brands in China mainly focus on five aspects: First, what kind of brand value should media organizations provide; Secondly, the differentiation of media products focuses on the positioning and personality style of media products; The third is the relationship between the audience and the brand; Fourth, media brand strategy research; Fifth, the relationship between media brand and media culture [5].

It can be found that those studies on brand issues are mostly on the theoretical level, and few of them targeted on enterprises. Studies on publishing media enterprises, brands and audiences are only limited to the discussion in between. There is almost no study on branding discussion by using short videos and other emerging media. Therefore, the establishment of brand communication influence evaluation model in publishing media industry plays a vital role in enterprise propaganda and content communication. There is no doubt that this paper is an action guide for publishing media enterprises to survive in the digital age.

3 Research Problems in This Paper

Based on the above studies and combined with the understanding of publishing and media enterprises' brands and brand media influence, this paper proposes the following four research questions:

- What are the communication factors to be evaluated based on the communication influence of publishing media?
- How to build an evaluation model based on the propagation influence of publishing media?
- How to verify the feasibility of the evaluation model for the propagation influence of publishing media?
- How to expand the brand influence of publishing media enterprises?

4 Propagation Breadth and Propagation Depth

4.1 Correlation Introduction

In the brand communication construction of publishing media enterprises centering on "single book (single set) brand, series brand, category book brand and publishing house brand", the leader brand, author brand and editor brand of publishing media enterprises should also be considered. We can all consider the brand communication of published content from the two perspectives of books and people.

Generally speaking, the propagation breadth of the short video refers to the number of time that short video content is forwarded on the media platform after the information content is published. The propagation depth of short video refers to the propagation influence of its stakeholders; That is, the information of a short video can be forwarded by subsequent users after being forwarded by an audience. The more individuals with propagation influence in short video users, the deeper the information will be spread. Based on the feature acquisition principle of mobile terminal devices in information dissemination, the author makes the following analysis according to the characteristics of brand communication of publishing and media enterprises.

Figure 2 is the horizontal lobe diagram of propagation depth. In the figure, range radiation from the center of the circle M at the divergence point and MP as the axis of symmetry shows that the target group can be precisely located while content

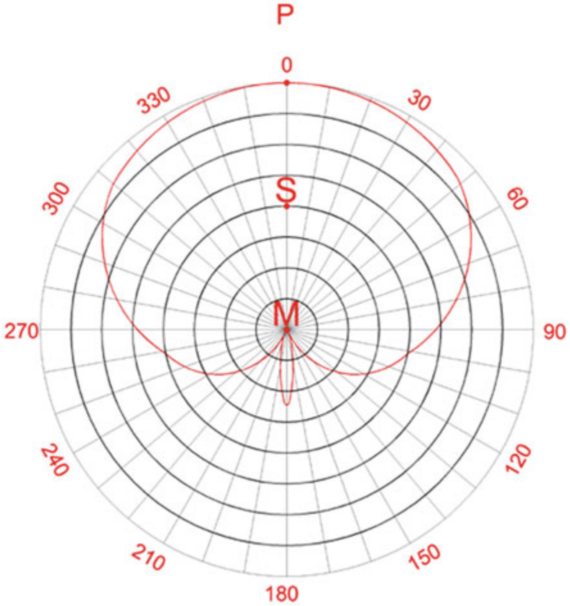
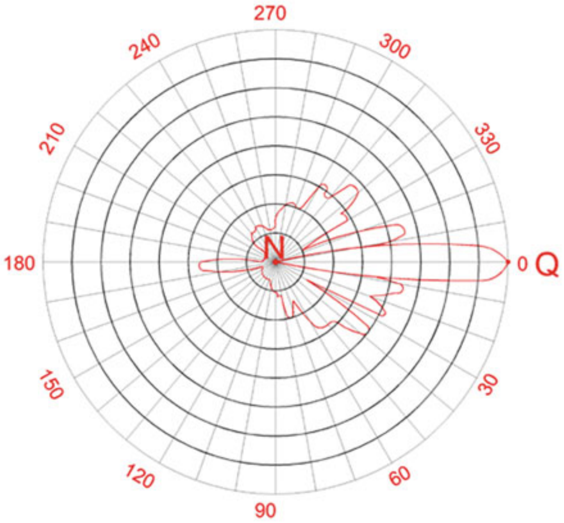


Fig. 2 Horizontal lobe diagram of propagation width

is the most significant. In the figure, range radiation centered on S is shown "0 degrees" radiates interested users without dead Angle. Through the channel of the short video, it innovates the way of publishing content telling, book discourse expression, and content communication form and communication channel and then realizes the 360-degree all-round communication pattern as showed in the figure through the forwarding of users. For example, when people's literature publishing house celebrated its 68th birthday in 2019, it invites translators such as Xu Yuanchong and writers such as Su Tong and Zhou Daxing to record a short video of their blessings, which gets close attention from readers. During the Spring Festival, Jie Li publishing house launched the "super flying man dream magic gift box" with a 40-second short video promotion in the store, selling 2000 sets in one week [6]. As the first publishing house to carry out self-publishing media business in China, *Xuelin Publishing House* broke through the time and geographical limitations by adopting the method of online public test on August 18, 2014 and presented authors, editors and books on the same platform, providing authors with two paths: Self-publishing and crowdfunding [7]. On December 21, 2018, *Xuelin Publishing House* released the first short video of TikTok, marking the official entry of *Xuelin Publishing House* into the short video category. The effective radiation of transmission width has realized the integrated development of publishing houses, users and technology companies, which made books and short videos work together, and effectively enhanced the brand influence of publishing and media industry.

Figure 3 is the vertical lobe diagram of propagation depth, which shows the "egg drop" propagation effect diagram with N as the original. When the content is concentrated in a certain vertical field for promotion, "Focusing the gathering, divergence

Fig. 3 Vertical lobe diagram of propagation depth

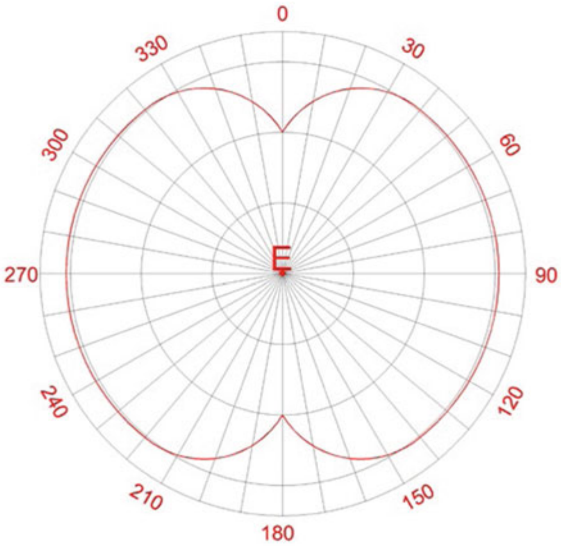


response" is the effect of propagation depth in publishing media industry, and finally reaches the furthest point Q of ideal propagation. When facing the red sea of the Internet, building a unique advantage through short videos is the secret to growing the brand tree in the field of publishing and media. Taking education publishing houses as an example, the activity of "the most beautiful children's voice" was carried out by *Chongqing University Press*, and the volume of thumb up reached 1.15 million. On November 8, 2018, *Double Eleven* Festival was taken into a short video content for the first time, promoting Lu Jingren's *Jingren language* and cartoon master Tango's and posters Jeams Jean's books. Famous foreign writer Yohji Yamamoto, Andrew Wilson's book section to make sound expression time attracted the attention of many users. Radiation of propagation depth realizes the user's drainage. Publishing house targeted services to users with the Internet user thinking, constructing user service platform and doing personalized brand communication. There is no doubt that it is the media publishing media industry in the depth of the effect that achieves the secret.

4.2 Effect

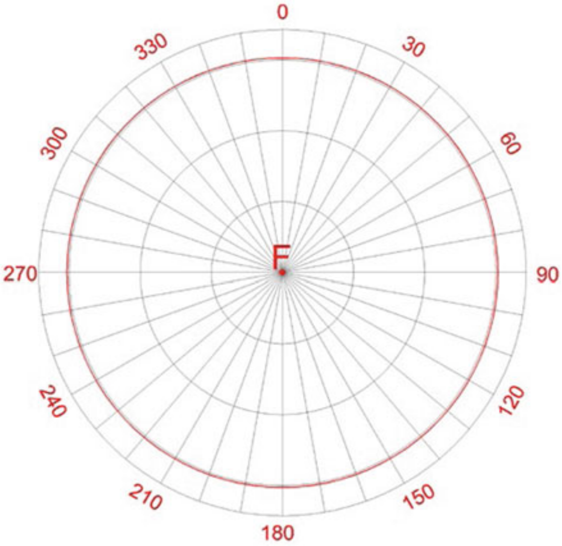
At the same time, we focus on the research content carried on the omni-directional collection: when a category is generally very big, we can observe it in a horizontal direction and vertical direction on covering observation, but the difference of horizontal and vertical direction can be concluded as: Due to the "propagation breadth", as defined by the group, the "blind area" is inevitable confusing, so data radiation pattern present fan-shaped petal shape as shown in Fig. 4. "propagation depth". As defined by the vertical field due to the characteristic, to conduct a comprehensive

Fig. 4 Influence radiation in horizontal direction



analysis, we can observe it in the control area because of its standardization, which is shown as Fig. 5. The omni-directional collection model can prepare for the operation of adequate data analysis. Accordingly it ensures the success of the operation and promotion.

Fig. 5 Influence radiation in the vertical direction



5 Evaluation Model of Brand Media Influence in Short Video

5.1 *Through Short Video to Discuss Brand Communication Elements*

Brand owners can conduct experiential marketing around each living consumer and implement the four laws of LOVE: Listen, Omini-channel, Value, and Engagement [8, 9]. And Online marketing is one of the best practices used to establish a brand and to increase its popularity [10]. The author believes that "LOVE" can be analyzed from the three elements of products, consumers and enterprises, as showed in Table 1.

It can be seen from Table 1 that products, consumers and enterprises are the basic elements for the construction of brand media influence. In the publishing and media industry, they are shown as: publishing short video content, short video users (readers), publishing and media enterprises. Combined with the communication characteristics of short video media, the author makes the following analysis.

As is shown in Fig. 6, a short video content is published in implementation to a short video as the core of brand communication. Through the social networking

Table 1 Dimension Division of Brand Communication and Shaping of Publishing Media Enterprises

The dimension	Consumer reference standard	Meaning	Publishing media enterprise brand building connotation	The essence of brand communication in publishing media enterprises
Product benefit points	Need	Why do consumers like it	Publications bring benefits to consumers	Functions and features of publications
Enterprise core value	Motivation	Why do consumers buy it	Brand core values for readers/viewers to bring life momentum	Reader's/viewer's driving influence
	Values	Why do consumers trust it	The strategic goal of brand building and development of publishing media enterprises	The embodiment of core competitive advantages of publishing and media enterprises
	Attitude	Why do consumers think highly of it		
Consumer's experience and feelings	Expection	Why do consumers buy it once again	The spirit of publications and the pursuit of ideas	Publishing media enterprise brand survival and development of the most valuable wealth

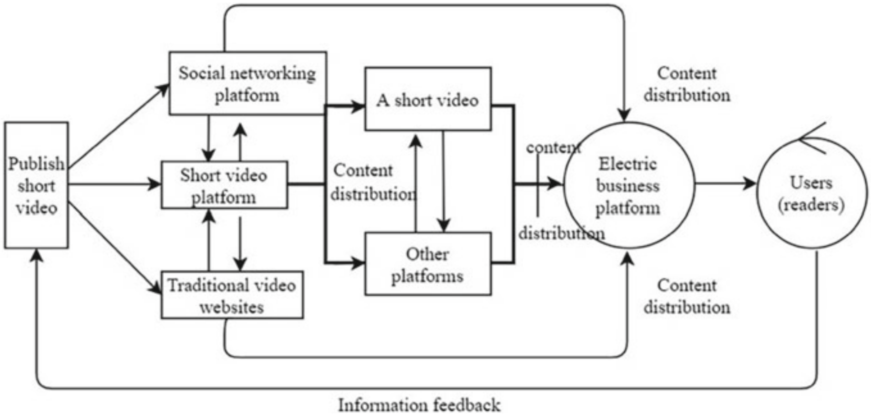


Fig. 6 Profit model of short video drainage publishing (reference from [11])

platform, the traditional video site causes flow. Multiple platforms cause content distribution. Finally, by the user (the reader) between the widespread propaganda and big data recommended, the publication medium enterprise realizes the popularization of short video content and promotes profit increase. Therefore, in the flow profit of short video publishing, we can use such a small-world, scale-free, self-organizing and clustered social network to measure the propagation breadth of the publishing media industry from two dimensions of contact influence and cognitive power. The depth of propaganda in the Publishing media industry is measured from satisfaction, loyalty and participation. See Fig. 7 for details.

Professor Yu Guomin pointed out that the essence of media influence is its own “channel brand” for social cognition, social judgment, social decision-making and related social behaviors of its audience as the channel of information dissemination. The most important thing in contact, maintenance and promotion is contacted and cognition [12]. Hence, it is necessary to use the characteristics of short video “short, flat and fast” to improve the audience’s access to the cognition of published media.

Satisfaction, loyalty and participation are important terms in marketing. In line with “technology first, the channel as the base, content as the most important status, doing everything for the user, service for more” [13], we should attach great importance to the user perception, media richness, social telepresence and social influence. In the sense of behavior loyalty and emotional loyalty [14] to intensify brand impact, we should shape the user in mobile scenarios with 4G for 5G network service upgrades pace to widespread media content.

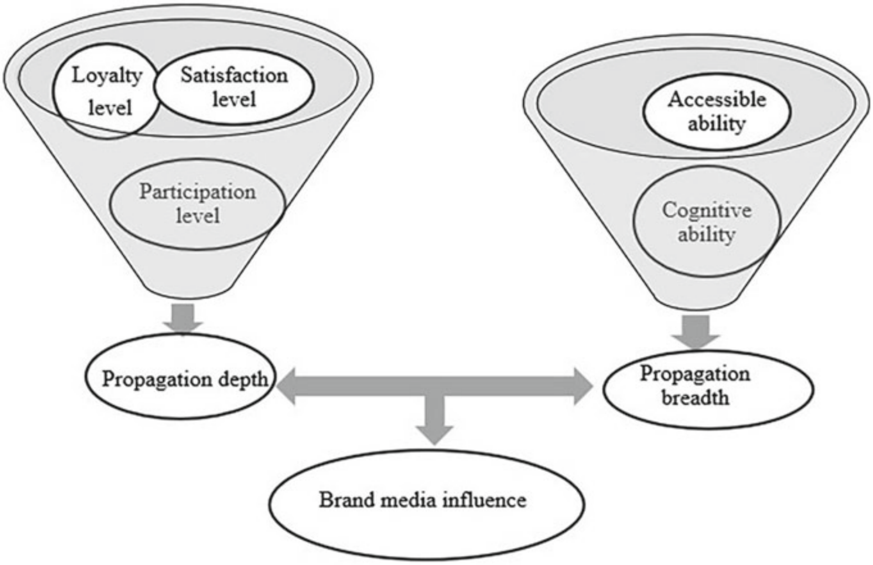


Fig. 7 Relationship diagram of components of brand communication influence

5.2 Evaluation Model Construction of Brand Media Influence in Short Video in Publishing Media Industry

1. *The index assumption of brand media influence evaluation model based on short video*

Zhang [15] from the perspective of "content influence, original influence, the influence of personality, sharing power, liquid influence" to analyze the spread of the media influence. Accordance with the final evaluation index quantification, the author combines with the contact of propagation characteristics of short video to conclude that quantitative indicators. That is Table 2.

2. *Quantitative evaluation assumption of brand media influences evaluation model based on short video*

Based on previous studies and the actual situation of short video transmission in the publishing and media industry, The author calculate formula is shown in Table 3.

Table 2 Table of publishing media industry communication influence quantification index

Evaluation content	The evaluation factors	Secondary index of evaluation	Note
Propagation breadth B	Accessible ability B_1	Number of users X_1	The number of users of publishing media enterprises
		User time X_2	Average user time per day
		Content quantity X_3	Total content by the end of the evaluation day
	Cognitive ability B_2	Content originality rate X_4	Content soundtracks accounted for a percentage of the full text during the evaluation period
		Focus on quantity X_5	Number of fans
Propagation depth D	Satisfaction level D1	Technical satisfaction X_6	Easy access to different terminal equipment, safe and reliable, beautiful design evaluation number
		Content satisfied X_7	Frequency of content updates
		Service satisfaction X_8	Feedback rate—ads break rate
	Loyalty level D_2	User viscosity X_9	Usage frequency during the evaluation period (default is the number of works)
		The recommended ratio is X_{10}	The registration rate recommended to other netizens
		Use intent X_{11}	Duration from after use to uninstall during evaluation period
	Participation level D_3	Thumb up number X_{12}	Thumb up during the evaluation period
		Comments number X_{13}	The number of comments during the evaluation period
		Forwarding number X_{14}	Forwarding during the evaluation period

Table 3 Table of quantitative calculation formulas

	Evaluation content	The evaluation factors	Secondary index of evaluation
Brand influence I (B + D)	Propagation breadth B ($B_1 + B_{2)}$	Accessible ability B1 ($X_1 + X_2 + X_3$)	User number X_1 (unit: 100 million)
			User usage time X_2 (unit: 100 million hours)
			Content quantity X_3
		Cognitive ability B_2 (X_5)	Content originality rate X_4
			Focus on quantity X_5
	Propagation depth D($D_1 + D_2 + D_3$)	Satisfaction level D_1 (X_7 * ($X_6 + X_8$)/$_2$)	Technical satisfaction X_6
			Content satisfied X_7
			Service satisfaction X_8
		Loyalty level D_2 (X_9 * $X_{10} + X_{11}$)	User viscosity X_9
			The recommended ratio X_{10}
			Use intention X_{11}
		Participation D_3 ($X_{12} + X_{13} + X_{14}$)	Thumb up number X_{12}
			Comments number X_{13}
			Forwarding number X_{14}

5.3 Verification and Analysis of the Evaluation Model of Brand Media Influence in Short Videos in Publishing Media Industry

Before the Spring Festival in 2019, there is an outbreak of COVID-19 in Hubei province. From January 24 (Lunar New Year's Eve) to February 8 (Lantern Festival), the author made the statistics from the entire publication medium enterprise. Taking TikTok short video application as an example, the author chose a new index of the top 10 list publication medium enterprise in NEWRANK website and the result is calculated by the following statistics (Tables 4 and 5).

According to the quantitative index of communication influence constructed above, the author made the following quantitative analysis on the top 10 publishing and media enterprises of the NEWRANK Index, and verified the feasibility of the quantitative table. The statistical table is shown in Tables 6 and 7.

Table 4 TikTok top 10 publishing media enterprise brand influence list during the spring festival

Ranking	Code	Name	Ownership	Publishing house (media company) category
1	①	Southern Plus Client	Southern newspaper media group	The official media
2	②	Live broadCast on nanyang cloud broadcasting station	Nanyang press media group Co. Ltd.	The official media
3	③	Nanyang newspaper media	Nanyang newspaper media	The official media
4	④	Huaihai evening	Huaihai evening news	Public publishing house
5	⑤	Nanyang daily	Nanyang daily	Public publishing house
6	⑥	Jia Shang media	Jia Shang media MCN	The private enterprise
7	⑦	A little book boy who can read	China academy of administration audiovisual publishing house	Professional publishing house
8	⑧	Leshan snail culture communication Co. Ltd.	Leshan snail culture communication Co. Ltd.	The private enterprise
9	⑨	Knowledge library	Mechanical industry press	Professional publishing house
10	⑩	Renjiao little Tik	People's education press	Educational publishing house

Table 5 Quantitative ranking list of brand influence of media enterprises published during the spring festival

Code	A_1^a	A_2^b (%)	A_3^c (%)	A_4^d (%)	A_5^e
①	872.8	87.28	99.58	99.37	1.8676
②	844.5	84.45	99.50	99.62	1.8401
③	818.9	81.89	98.76	98.10	1.8032
④	796.6	79.66	97	98.42	1.776
⑤	778.4	77.84	98.60	97.64	1.7596
⑥	769.6	76.96	98.27	98.78	1.7549
⑦	416	41.60	79.27	85.72	1.241
⑧	541.9	54.19	82.08	5.26	0.9786
⑨	348.1	34.81	29.58	37.92	0.6856
⑩	407	40.70	25.01	25.67	0.6604

[a] NEWRANK index

[b] NEWRANK index unitized: (a1/1000)

[c] TikTok propagation influence

[d] TikTok influence

[e] Comprehensive brand influence: ((A3 + A4)/2 + A2)

Table 6 Statistical table of quantitative analysis of publishing media enterprises quantitative ranking list of brand influence of media enterprises published during the spring festival

Content	Factors	Secondary index	①	②	③	④	⑤
B	B_1	X_1	4	4	4	4	4
		X_2[a]	4.65986	4.65986	4.65986	4.65986	4.65986
		X_3	704	659	493	202	485
	B_2	X_4	44.71%	100%	100%	66.67%	63.94%
		X_5	800797	660,506	169,158	270,473	127,699
D	D_1	X_6	100%	100%	100%	100%	100%
		X_7	85	228	159	51	208
		X_8	100%	100%	100%	100%	100%
	D_2	X_9	85	228	159	51	208
		X_{10}	100%	100%	100%	100%	100%
		X_{11}	16	16	16	16	16
	D_3	X_{12}	8,171,226	7,019,176	2,960,657	1,262,576	1,233,251
		X_{13}	435,549	203,656	2048	5125	2053
		X_{14}	320,900	3,484,458	22,232	257,091	34,004

Table 7 Statistical table of quantitative analysis of publishing media enterprises

Content	Factors	Secondary index	⑥	⑦	⑧	⑨	⑩
B	B_1	X_1	4	4	4	4	4
		X_2[a]	4.5986	4.5986	4.5986	4.5986	4.5986
		X_3	238	291	466	117	80
	B_2	X_4	100%	100%	100%	0	100%
		X_5	1,609,012	167,019	937,661	84,464	95,978
D	D_1	X_6	100%	100%	100%	100%	100%
		X_7	6	6	12	1	1
		X_8	100%	100%	100%	100%	100%
	D_2	X_9	6	6	12	1	1
		X_{10}	100%	100%	100%	100%	100%
		X_{11}	16	16	16	16	16
	D_3	X_{12}	850,086	2128	44,700	−492.	−837.
		X_{13}	7408	7	586	0	11
		X_{14}	1274	53	88	1	9

[a] Explanation: according to data from *Yiguan Qianfan*, during the COVID-19 period, the usage time of TikTok increased to 459.86 million hours during the five days from January 25 to 30, both of which reached the best performance in January [16]. Moreover, the number of daily active users of TikTok has exceeded 400 million as of February 8, which is taken as a reference in this paper [17]

Table 8 Table of statistical results of quantitative analysis

R_1	R_2	C	I	B	D	B1	B2	D1	D2	D3
1	2	①	11,368,927	661,165	10,707,762	659	660,506	228	244	10,707,290
2	1	②	729,362	801,501	8,927,861	704	800,797	85	101	8,927,675
3	3	③	3,154,922	169,651	2,985,271	493	169,158	159	175	2,984,937
4	6	④	2,468,046	1,609,250	858,796	238	1,609,012	6	22	858,768
5	4	⑤	1,795,585	270,675	1,524,910	202	270,473	51	67	1,524,792
6	5	⑥	1,397,924	128,184	1,269,740	485	127,699	208	224	1,269,308
7	8	⑦	983,541	938,127	45,414	466	937,661	12	28	45,374
8	7	⑧	169,526	167,310	2216	291	167,019	6	22	2188
9	10	⑨	95,259	96,058	−799	80	95,978	1	17	−817
10	9	⑩	84,108	84,581	−473	117	84,464	1	17	−491

Description: X_1 and X_2 are all data of TikTok data platform, which are ignored when the size is larger

R_1: Current ranking

R_2: NEWRANK quantifies rankings

C: Code

According to the data in the quantitative statistics table and the calculation formula in Tables 6 and 7, the following calculation results are obtained.

As can be seen from Table 8, compared with the quantitative ranking of NEWRANK, the statistical results show a cluster distribution, which is 1–3, 4–6, 7–8, and 9–10, respectively. Each TikTok sign is kept in the original cluster interval, and it is floating around 1 rank. It is worth noting that *Jia Shang media* did not show any work on the COVID-19 during the Spring Festival, and its statistical ranking rose two places, presenting a special situation.

The above verification analysis shows that the evaluation model of the brand media influence of publishing media in short videos is feasible and can guarantee the stability of the interval.

6 Conclusions

By constructing an evaluation model of the brand media influence of publishing media in short videos, this paper studies the brand media influence of publishing media enterprises in short videos, and attempts to construct a quantitative evaluation index for the communication influence of publishing media enterprises based on the characteristic factors of brand media influence. Based on the research results of this paper, the following suggestions are proposed to expand the brand influence of publishing and media enterprises:

1. *Pay attention to the construction of brand ecological environment and let users benefit from the premium content and technology*

Through the research, it can be found that the publishing and media enterprises do not make full use of resources for brand promotion, resulting in a great waste of resources; On the technical level, small independent innovation ability is the biggest problem to restrict publishing and media enterprises, which shows the importance of applying block chain, big data and intelligent contract to copyright asset management. The construction of brand influence cannot be separated from users.

2. *Make sure content quantity and quality development at the same time*

The amount of content is the most direct response to the communication strength of publishing and media enterprises, which will lead to the increase of the number of users thumb up and the number of comments. But there still might be some publication medium enterprise pursuit one-sided effect to expand influence, ignoring the quality of content. It may result in the "deformed triangle" of resources, technology and human rights on the road to achieve consistent development.

3. *Put the block chain technology into practice and run it through every link of brand communication*

For a long time, Chinese enterprises rely on the advantages of cost, labor force, resources and environment and the introduction of technology for local regional development [18]. In the research process, the author finds that this is particularly evident in publishing and media enterprises. As for content providers, publishing media companies, platform operators, and service providers, they need to pay attention to technology development at any link. It's not negligible to make full use of chain blocks technical features, which will be open and transparent to everyone so as to participate in the database records. It is the foundation to combine iterative development and application ecology of block chain technology with brand communication provide users with high quality and guarantee.

The deficiency of this paper lies in that the quantity index and the calculation formula of the evaluation model of brand communication influence are based on the summary of previous studies and lack of verification analysis. There is no comparative analysis of the evaluation model and no comparative analysis of the model's advantages. The amount of research data is limited to the statistical results during the Spring Festival in 2020. This paper is preliminary experimental results of brand research, which is still in progress. However, it can be seen that the construction of this model fills the blank of the evaluation system of brand communication influence of publishing media enterprises. With the horizontal lobe diagram and the influence radiation diagram, this paper expounds the idea of communication breadth and communication depth. Using the novel Coronavirus content data of publishing media enterprises during the outbreak of COVID-19 at the beginning of 2020 can provide reference for exploring the way and method of brand communication of enterprises in the special period.

Acknowledgements This work is partially supported by the Key Program of Beijing Social Science Fund "Research on the construction and application of the ecosystem of Beijing digital copyright industry based on blockchain" under contract with No. 19JDXCA001.

References

1. Chinese publishing innovation conference held in Beijing in 2020. [EB/OL]. http://www.chuban.cc/zx/cbyw/202001/t20200114_181643.html. Publishing media industry in China. January 8, 2020
2. N. Hollis, How brands influence purchase decisions, in *Brand Premium* (Palgrave Macmillan, New York, 2013), pp. 25–37
3. X. Chang, J. Yuwen, The development direction of reference book publishing in China from the current situation of foreign publishing industry in the digital era. Libr. J. **38**(09), 62–66 (2016)
4. X. Lu, Z. Lu, Fifteen seconds of fame: a qualitative study of Douyin, a short video sharing mobile application in China, in *Social Computing and Social Media. Design, Human Behavior and Analytics*. Lecture Notes in Computer Science, HCII 2019, vol. 11578, ed. by G. Meiselwitz (Springer, Cham, 2019), pp. 233–244
5. Zhou Xue, Research on overseas dissemination of Tiktok short video and self-modeling of national image. Audiovisual **09**, 151–152 (2019)
6. Chinese publishing innovation conference held in Beijing in 2020. [EB/OL]. http://www.chuban.cc/zx/cbyw/202001/t20200114_181643.html Publishing Media Industry in China (2020)
7. In the second half of the short video, how did the press break the game? The Chinese press and publication telegraph http://www.xinhuanet.com/zgjx/2019-04/15/c_137978267.htm (2019)
8. Y. Xiang, Research on brand marketing of mobile short video. Hunan Normal University (2019)
9. N. Zhu, Research on digital development of publishing and media enterprises. Huazhong University of Science and Technology (2012)
10. N. Aggrawal, A. Ahluwalia, P. Khurana, A. Arora, Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media. Soc. Netw. Anal. Min. **7**(1) (2017)
11. Chen Juhong, Research on short video marketing in publishing media industry in the era of mobile Internet. Publ. Sci. **27**(04), 80–84 (2019)
12. Yu. Guoming, Interpretation of media influence—a discussion on the nature of media industry. Int. Press **02**, 5–11 (2003)
13. Zheng Baowei, Current status and characteristics of journalism research in China. New Media Soc. **1**, 45–49 (2015)
14. Liu Hongcheng, Integration of customer loyalty and customer relationship management [J]. Contemp. Econ. **08**, 26–27 (2004)
15. Z. Haitao, Z. Huiran, W. Ping, L. Tianyin, Evaluation model construction of new media information transmission influence from the perspective of super IP. Intell. Sci. **37**(02), 3–8 (2019)
16. The longest in the history of the Spring Festival holiday “boring books:” peace elite squeezed collapse trill quickly swelled hours used 100 million hours [EB/OL]. https://baijiahao.baidu.com/s?Id=1657681796530693462&WFR=spider&for=PCflushfinance, 5 Feb 2020
17. 2019 trill data report (full version). [EB/OL]. https://mp.weixin.qq.com/s/a5pYni2h5AKKgqp0vQsEAA.CEOgen, 5 Feb 2020
18. H. Jingdong, H. Qunhui, Factor analysis of influencing industrial service outsourcing—panel data analysis based on 22 industrial industries. China Ind. Econ. **12**, 44–56 (2012)

Balancing Optimization of Mixed-Flow Assembly Line Based on Hybrid Genetic Algorithm

Meng Li and Dan Chang

Abstract The balancing problem of the assembly line is a typical NP-hard problem and one of the most important problems to be solved by manufacturing companies. For the disadvantages that genetic algorithms are prone to local optimal solutions and precociousness in the optimization process, the article focuses on the second type of balancing issues for mixed-flow assembly lines. A hybrid genetic algorithm was designed and constructed to combine three evaluation metrics: the smoothing index, the equilibrium loss coefficient and the imbalance coefficient of the adaptation function, combining the simulated annealing algorithm with a genetic algorithm to speed up convergence to obtain a global optimal solution. Finally, a mixed-flow assembly line of Company L is used as an example to solve the equilibrium of the assembly line using the designed hybrid genetic algorithm for improvement and optimization. And also demonstrates that the method is suitable for solving balance problems in mixed-flow assembly lines.

Keywords Hybrid genetic algorithm · Simulated annealing algorithm · Mixed flow assembly line balancing

1 Introduction

In a mixed-flow assembly line, different types of products are often assembled simultaneously in the production line, depending on the requirements of different production tasks. The related products are similar in structure and process during the assembly process. A mixed-flow assembly line can respond quickly to market demands for different product types without taking up large amounts of inventory. The assembly line balancing analysis means that the different assembly tasks for different products are assigned to each station by means of optimization analysis and

M. Li (✉) · D. Chang
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 19125494@bjtu.edu.cn

D. Chang
e-mail: dchang@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_64

calculation. Making the operating time of a station less than or equal to a predetermined cycle time is a typical combinatorial optimization problem in engineering [1]. The assembly tasks and times for different products in a mixed-flow assembly line are often different, which poses a great challenge for balancing tasks.

In this paper, a hybrid genetic algorithm is designed for the Type II balancing problem in mixed-flow assembly lines, considering the combination of smoothing exponential, equilibrium the adaptation function for the three evaluation indicators of loss coefficient and imbalance coefficient combines the simulated annealing algorithm with genetic algorithm(GA) to overcome the traditional the slow convergence of the genetic algorithm and the tendency of the simulated annealing algorithm to fall into local optimal solutions [2] in order to speed up the convergence rate to obtain the global optimal solution is solved by simulation using MATLAB (2018a). Finally, using an L company's instrument assembly line as an example, the equilibrium model is used to optimize the equilibrium of this mixed-flow assembly line, and the cycle time of the assembly line number improved from 219 to 195 s; equilibrium index changed from 34.8 to 7.87, an improvement of 26.93; equilibrium loss the coefficient changed from 13.6 to 3%, an improvement of 10.6%, effectively solving the balance of the company's instrument assembly line, while the feasibility of the hybrid genetic algorithm is demonstrated. Comparing the improvement results of the hybrid genetic algorithm incorporating simulated annealing with those of the traditional genetic algorithm, it is possible to obtain a hybrid the faster convergence of the genetic algorithm proves that the method is suitable for solving the equilibrium problem of the mixed-flow assembly line.

2 Literature Review

In our current low level of manufacturing, many manufacturing companies still have backlogs of items due to assembly line balancing problems, assembly line low efficiency and high production costs. Especially for mixed-flow assembly lines, such problems are more likely to occur. Therefore, scholars have conducted extensive research and proposed different solutions for such problems.

Linear and dynamic planning are the most widely used methods for solving assembly line balancing problems. A 0–1 integer planning model for MMALBP problem characteristics. Scholl and Boysen [3] designed the dynamic planning model for parallel assembly lines. Li [4] obtained the optimal operating solution for the process in Company H through integer planning model. However, this method is difficult to apply to actual production for the balance of mixed-flow assembly lines with large structure and scale.

The assembly line balancing related technology mainly based on Industrial Engineering (IE) method is mainly two basic technologies of operation method research and operation measurement. He and Zheng [5] analyzed the PC machine assembly line of Company B, model method and 5W2H method to improve the bottleneck process. For the balancing of mixed-flow assembly lines alone, industrial engineering methods are generally used for the initial balancing of the assembly line, and optimization is mostly based on the theoretical basis, the actual effect can only be obtained when the optimization is completed, and to further improve the equilibrium effect, artificial intelligence is generally used algorithm.

Artificial intelligence algorithm-based production line balance optimization is a more efficient production line balance optimization method, which has been studied by many scholars in recent years direction. Zhao [6] studied the production balance problem and applied a genetic algorithm to improve the production line to obtain the minimum number of workstations; Zhang [7] applied genetic algorithms to solve the Type II production balancing problem; Fang [8] studied the production line balancing problem in order to Minimizing the number of production line workstations and minimum load smoothing index are the goals for production line optimization. Genetic algorithms are widely used in the balancing problem of the assembly line, more and other algorithms are combined to form new intelligent algorithms to overcome the genetic algorithm inherent flaws. The genetic algorithm is slow to converge due to its strong global search capability and poor local search capability, while the simulated annealing algorithm has the strong local search capabilities, and thus, incorporating simulated annealing algorithms into traditional genetic algorithms, combined into hybrid genetic algorithms. It compensates for the shortcomings of traditional genetic algorithms and has certain advantages.

Although planning is the most commonly used method for solving assembly line balancing problems, it is not suitable for mixing flows of larger structures and sizes. Assembly line balancing, difficult to apply to actual production. Although industrial engineering methods are well established, they are generally used for initial balancing of assembly lines, and the improvement process is too dependent on theoretical evidence. The optimization is completed before the realization effect can be obtained, and to further improve the improvement effect, artificial intelligence algorithm is generally used. Genetic algorithms are widely used in the balancing problem of assembly lines, but they are characterized by strong global search ability and poor local search ability. Therefore, its convergence rate is slow. The simulated annealing algorithm, on the other hand, has a strong local search capability, so it is important to incorporate the simulated annealing algorithm into the traditional genetic algorithm in combination with the become a hybrid genetic algorithm to compensate for the shortcomings of traditional genetic algorithms. Therefore, the hybrid genetic algorithm that combines genetic algorithm with simulated annealing algorithm in the study of equilibrium problems in mixed flow assembly lines has the superiority.

3 Analysis Review of Balance Problems in Mixed-Flow Assembly

The balance of the assembly line is a common problem faced by manufacturing companies, so the balance of the assembly line is one of the most important goals pursued by the company. In contrast to a single line such as a single product, a mixed-flow assembly line allows products with different processes and operations to be assembled on the same line. For production assembly, with a greater emphasis on real-time production and product variation, which increases the complexity of production management to some extent. It becomes a difficult issue to study.

3.1 Existing Cycle Time Does not Meet Consumer Demand

Therefore, further research on the balance of the mixed-flow assembly line is needed to reduce the cycle time of the assembly line, speed up the production efficiency of the enterprise, and meet the market demand for products.

3.2 Uneven Load Distribution Among Workstations on the Assembly Line

Due to the large number of products produced on the mixed-flow assembly line, the assembly process may not be identical between different products, which often leads to some processes being idle and some processes being too busy, causing the psychological imbalance between workers to rise, greatly affecting the enthusiasm of the workers, reducing the efficiency of the assembly and indirectly leading to a decrease in the quality of the product assembly [9].

The equilibrium loss factor, also known as equilibrium delay, reflects the extent to which unbalanced idle time affects assembly line ineffectiveness, determining whether the assembly line is load-balanced is one of the important criteria for evaluating the balance of the production line. The size of the balance loss coefficient determines the degree of balance of the assembly line, see Table 1.

Therefore, it is necessary to the optimization of the mixed-flow assembly line and to evaluate the load balance of the assembly line by the equilibrium loss factor.

Table 1 Criteria for judging the balance of the assembly line

Equilibrium loss factor	Assembly line judgment results
$d \le 10\%$	Excellent
$10\% < d < 20\%$	General
$10\% < d < 20\%$	Poor

3.3 Complexity of Mixed-Flow Assembly Lines

Since the logical order between different products on a mixed-flow assembly line is not necessarily the same, in the study of multi-species mixed-flow assembly, the processing sequence logic diagrams and operational tasks of these products are generally combined together and merged into the same sequence logic diagram to form a unified multi-product assembly task, see Figs. 1 and 2.

From the figure above, it can be seen that the mixed-flow assembly line is a multi-species mixed-flow assembly line, in which three products can be assembled simultaneously on one assembly line. Compared to single-species assembly, the mixed-flow assembly line combines the assembly tasks of different products and the processing sequence diagram. The logic diagram will be more complex and the optimization process will be more complicated [10]. For the optimization study of complex mixed-flow assembly line balancing problems, artificial intelligence algorithms are widely used and have advantages, so the use of the algorithm is superior for balancing optimization of mixed-flow assembly lines.

In this paper, we design a hybrid genetic algorithm that integrates genetic algorithms with simulated annealing for the Type II equilibrium problem of mixed-flow assembly lines. Taking an instrument assembly line of Company L as an example, the

Fig. 1 Assembly sequence logic

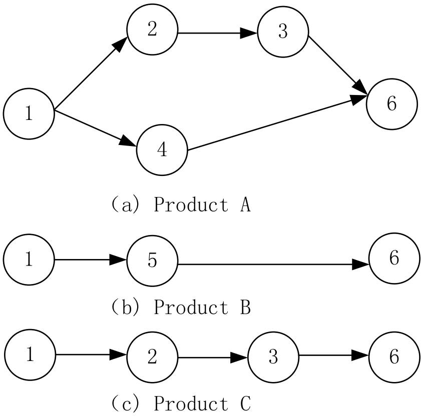
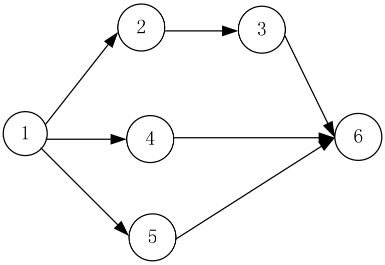


Fig. 2 Post-merger assembly sequence logic diagram



balance model is used to optimize the balance of the mixed-flow assembly line and the cycle time of the assembly line. It is optimized to solve the balancing problem in the assembly line of this factory and to prove the feasibility of the hybrid genetic algorithm. Comparing the improved results of the hybrid genetic algorithm incorporating simulated annealing with the improved results of the traditional genetic algorithm, it is possible to obtain a hybrid genetic algorithm. The faster convergence of the hybrid genetic algorithm proves that the method is suitable for solving the equilibrium problem in mixed-flow assembly lines and it has certain advantages.

4 Equilibrium Model Construction Based on Hybrid Genetic Algorithm

Genetic algorithms and simulated annealing algorithms are highly complementary; genetic algorithms are more capable of global search, but tend to fall into the local optimum, appearing precocious, while the simulated annealing algorithm has a strong local search capability and can effectively avoid the local optimum, but its global search capability is limited. By mixing the two algorithms, it is possible to overcome their respective shortcomings and balance the algorithm's ability to search intensively and extensively, thus achieve better optimization results [11].

4.1 Hybrid Genetic Algorithm Design

The balancing problem of an assembly line is a typical NP-hard problem, and this paper adopts a hybrid genetic algorithm to solve this problem. An analysis is conducted to combine the genetic algorithm with the simulated annealing algorithm to effectively improve the search potential of the genetic algorithm. The structure of the hybrid genetic algorithm combined with the simulated annealing algorithm is shown in Fig. 3.

1. *Basic Flow of Hybrid Genetic Algorithms*

(a) *Coding design*: For different problems, the same coding rules are used, and their computational performance for solution varies. In this paper, we study NP problems such as mixed-flow assembly line balancing, and specifically design their specific chromosome coding rules - based on operation sequential chromosome coding rules.

(b) *Initialization of the population*: For the mixed-flow assembly line equilibrium problem to be investigated in this paper, the generation of the initial population will be constrained by the order of operations, which must meet the requirement of randomness and not violate the order of operations constraint, so this paper is designed on the premise of satisfying the operation priority relationship, firstly

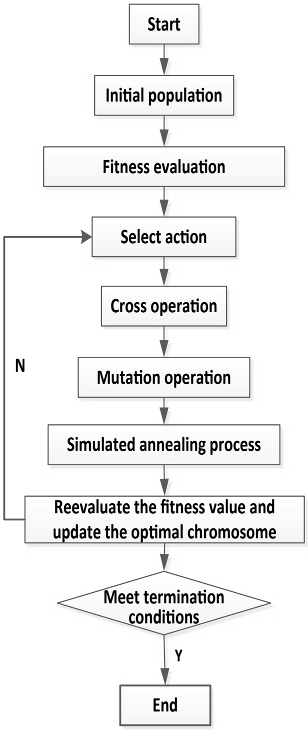

Start
Initial population
Fitness evaluation
Select action
Cross operation
Mutation operation
Simulated annealing process
Reevaluate the fitness value and update the optimal chromosome
Meet termination conditions
N
Y
End

Fig. 3 Structural diagram of the hybrid genetic algorithm

to generate an alternative set that conforms to the priority relationship, and then to randomly select an element from the alternative set to be programmed into the chromosomal locus, until all the operation elements are selected into the locus, then a feasible numerical sequence can be generated and used as an individual of the initial population.

(c) *The genetic operator*:

- Selection. Assume that the population size is n, the value of the fitness function for individual i is $f(U_i)$, and the sum of the fitness values of all individuals within the population is F, according to roulette selection, it is known that individual i is selected with probability $P_i = f(U_i)\Big/ F$ in the calculation. therefore, with probability P_i for the population of choice.
- Crossover. The biggest difficulty in the design of the crossover operation in this paper is that it is constrained by the operational priority relationship, which must ensure that individual chromosomes remain after the operation conform to the priority relation, otherwise it will be a non-feasible solution, then the operation will not proceed.
- Variation. In order to expand the search space of the genetic algorithm and avoid getting local optimal solutions, the variation operation is needed. Due to the restriction of the operational processing sequence in the production line,

the mutated individuals may not meet the actual production requirements, so it is necessary to perform the mutated individuals are tested to determine if they satisfy the logical relation of the priority relationship matrix and only if the new individual with the variation is superior to the old individual, in order to renew it as a new individual.

(d) *Simulated annealing process*: In order to overcome the limitations of the traditional single genetic algorithm based on the roulette idea [12], many authors have separately incorporated other intelligent algorithms of factors introduced into genetic algorithms, resulting in many new intelligent combinatorial optimization algorithms. In essence, these new optimization strategies for simulating natural phenomena are added to the selection process of genetic algorithms in order to solve the problems in the problem of offspring "freshness" in the process of parental selection and offspring generation. Jumping out of the "trap" of local optimum solutions to reach a global or engineering optimum on the premise of a relatively limited population [13]. In this paper, the idea of solving the problem of precocious maturation of genetic algorithms is to introduce the strategy of "genetic annealing" [14, 15]. The solution to the problem of precocity of genetic algorithms in this paper is to introduce the strategy of "genetic annealing".

2. *Basic Flow of Hybrid Genetic Algorithms*

(a) *Smoothing index*: In a balanced assembly line, for the second type of balancing problem, the aim is to reduce the cycle time with the number of stations known, but while pursuing the minimum cycle time, the load balance between the various stations on the assembly line is required, expressed by the smoothing index SI of the line, with a smaller SI indicating a more balanced load at the line stations.

$$\mathrm{SI} = \sqrt{\frac{\sum_{i=1}^{N} (CT - T_i)^2}{N}}$$

Where CT is the cycle time, which is the processing time per station, and N is the number of assembly line stations.

(b) *Equilibrium loss factor*: A well-balanced assembly line requires not only load balancing between assembly line stations, but also the reduction of cost factors and loads [16–18]. Here an equilibrium loss factor d is introduced.

$$\mathrm{d} = \frac{N * \max(T_i) - \sum_{i=1}^{N} T_i}{N * \max(T_i)}$$

(c) *Imbalance coefficient*: For the topic of Type II balancing on assembly lines, evaluating the merits of an assembly line balancing design solely on the basis of balancing losses is sometimes not comprehensive enough, because if the balance loss of the assembly line is the same, the operating time balance of the assembly workstations may sometimes be very different [19–22]. Therefore, this paper introduces another assessment of the merits of the assembly line balance indicators - imbalance coefficient. The imbalance coefficient can be defined as the difference between the maximum and minimum workstation operating time and the ratio of the average workstation operating time.

$$\mathrm{k} = \frac{\max(T_i) - \min(T_i)}{ave(T_i)}$$

where: $\max(T_i)$, $\min(T_i)$ and $ave(T_i)$ denote the maximum workstation operating time, minimum operating time and average workstation operating time, respectively. Obviously, the smaller the imbalance coefficient, the better the assembly line balance for the same balance loss.

(d) *Adaptation function*: All assembly units need to be assigned to the appropriate workstations, minimizing the number of workstations and equalizing the load between workstations for assembly line production balance. Considering the workstation losses, equilibrium loss condition and equilibrium condition of the assembly line, the adaptation function is constructed as follows.

$$f = \omega_1 \frac{1}{SI + 0.01} + \omega_2(1 - d) + \omega_3(1 - k)$$

where ω_1, ω_2 and ω_3 are the coefficients of the weights. The first part of the function ensures the balance of the assembly line, and the second part ensures minimal losses and minimal cycle time at each station in production. Part 3 guarantees the degree of workstation operation time balance, generally taking $\omega_1 + \omega_2 + \omega_3 = 1$. The values of ω_1, ω_2 and ω_3 are based on the objective function on the assembly the balance of the production line and the number of workstations are dependent on the requirements. The fitness function is used to measure the adaptability of each individual in a population, and individuals with higher fitness have a higher probability of being passed on to the next gene.

4.2 Model Effect Validation

In this paper, the validity of the model is verified using the classical Jackson production line balancing problem, Jackson uses the enumeration method to optimize the balancing of the production line, Jackson's process priority diagram is shown in Fig. 4, and the time of each process is Time = [6 2 5 7 1 2 3 6 5 4].

Balanced optimization of the production line using a hybrid genetic algorithm based on a matrix of time and optimization relationships for each stage of the Jackson line. The optimized sequence of each workstation and process is shown in Fig. 5. The Jackson line balancing problem each station time, the line equilibrium rate and equilibrium index and the time per station, line equilibrium rate and smoothing index using the hybrid genetic algorithm are shown in Table 2.

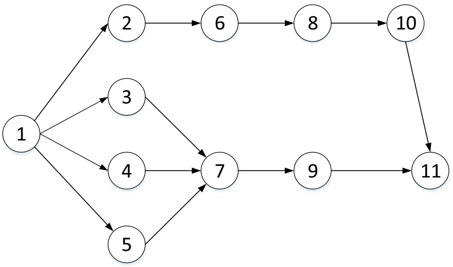


Fig. 4 Jackson process prioritization chart

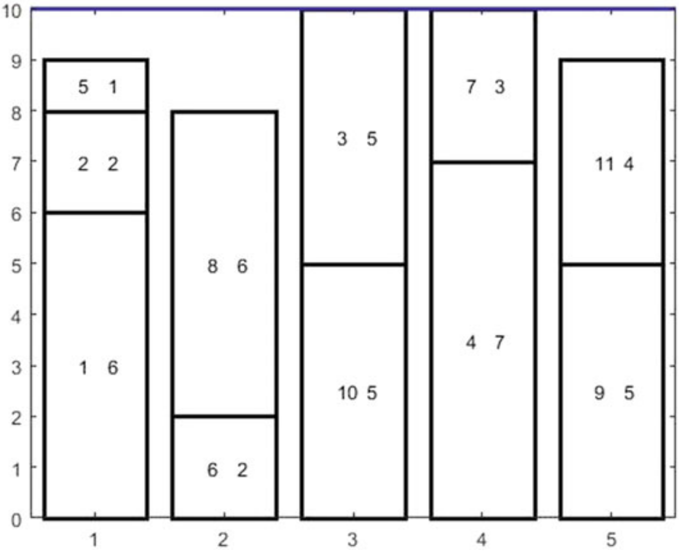


Fig. 5 Jackson production balance scenario

Table 2 Jackson production balance indicators

Stage	Workstation number					Production balance indicators			
	1	2	3	4	5	SI	d	k	Objective function
Current situation	10	7	10	10	9	3.16	92%	0.33	0.35
Genetic algorithm	9	8	10	10	9	2.50	92%	0.22	0.67

From Table 2, it can be seen that the equilibrium loss factor after balancing the production line by the genetic algorithm and the Jackson enumeration method is the same, but the smoothing index is improved, and the smoothing index of the production line balanced using the hybrid genetic algorithm is less than the Jackson status quo, thus indicating that the hybrid genetic algorithm is effective in optimizing the balancing of the assembly line.

5 Case Calculations and Analysis

5.1 Analysis of the Current State of Assembly in Company L

Take the instrument assembly line of an L company as an example, its mixed assembly line first assembles two kinds of experimental apparatus, namely, light chaser A and winning amusement machine B. The two products are produced on a fixed ratio of mixed lines. The integrated assembly ratio of product A and product B is 1:1. According to the mixing ratio of the two products. Prepare a joint operation priority relationship diagram for the two products, see Fig. 6, where the combined operation time for products A and B is 30 s. The two products will be used as a single product.

Company L's instrument assembly line is a mixed-flow assembly line with a total of nine workstations, with the distribution of processes for each workstation as shown in Table 3.

From Table 3, it can be seen that Station 5 has the longest operating time, 221 s, and the assembly line has a smoothing index SI of 36.5. The equilibrium loss coefficient

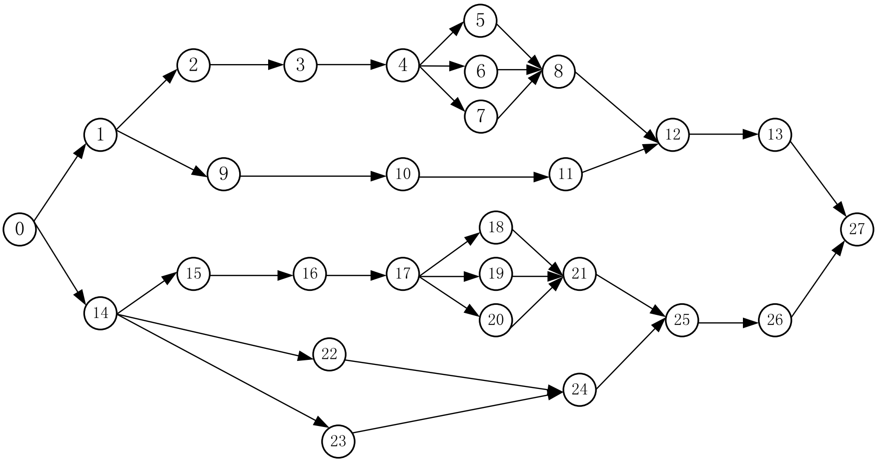


Fig. 6 Relationship between priorities for joint operations of two products

Table 3 Current plant distribution scheme

Work station	Process	Time	Work station	Process	Time
1	14, 15	176	6	5, 6, 7, 10, 11, 24	175
2	16, 23	177	7	8, 12, 13	185
3	1, 2	176	8	17, 18, 19, 20	197
4	9, 22	173	9	21, 25, 26, 27	219
5	3, 4	221			

is 14.4%, greater than 10%. The assembly line is in the "good" category and needs to be further improved and optimized.

5.2 *Assembly Line Balance Optimization*

Using MATLAB (2018a) simulation software to optimize the balance of this mixed-flow assembly line using a hybrid genetic algorithm. Where the number of population iterations is 100, the crossover probability is 0.6, the mutation probability is 0.15, and the temperature reduction parameter c = 0.98. The weighting coefficients, and sums of the adaptation function are 0.4, 0.3, and 0.3, respectively. After 100 iterations, the optimization results are shown in Fig. 7. The distribution of work processes in the workstations tends to be balanced, and the cycle times are optimized. Nine workstations are distributed as shown in Table 4.

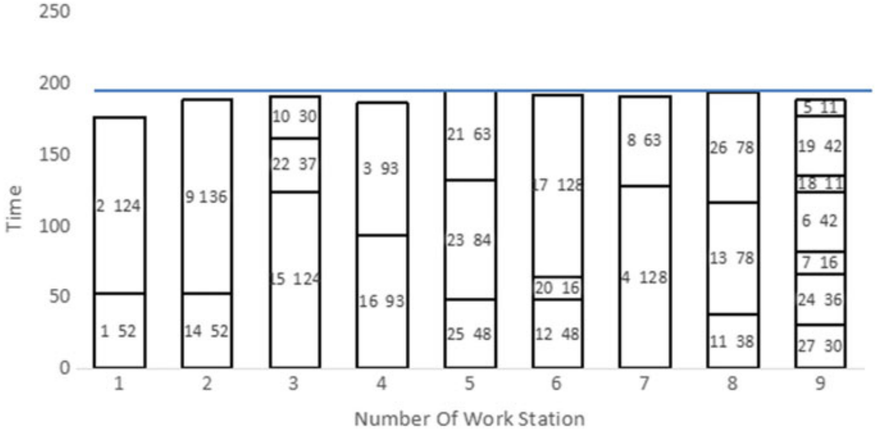


Fig. 7 Process distribution chart

Table 4 Improved plant workstation assignment programme

Work station	Process	Time	Work station	Process	Time
1	1, 2	176	6	12, 17, 20	192
2	9, 14	188	7	4, 8	191
3	10, 15, 22	191	8	11, 13, 26	194
4	3, 16	186	9	5, 6, 7, 18, 19, 24, 27	190
5	21, 23, 25	195			

5.3 Analysis of Results

1. *Improvement of Impact Evaluation*

Comparing the before and after improvements to Company L's instrument assembly line, the process assignments and station cumulative times for the nine workstations are shown in Table 5. The distribution of work processes at each station is balanced, and the comparison of evaluation indexes before and after improvement is shown in Table 6. The value of cycle time on the assembly line improved from 221 to 195. The balance index changed from 36.53 to 7.87, reducing the value of cycle time 28.66.

Table 5 Comparison of process distribution before and after improvement

Work station	Pre-improvement		Improved	
	Process	Time	Process	Time
1	14, 15	176	1, 2	176
2	16, 23	177	9, 14	188
3	1, 2	176	10, 15, 22	191
4	9, 22	173	3, 16	186
5	3, 4	221	21, 23, 25	195
6	5, 6, 7, 10, 11, 24	175	12, 17, 20	192
7	8, 12, 13	189	4, 8	191
8	17, 18, 19, 20	197	11, 13, 26	194
9	21, 25, 26, 27	219	5, 6, 7, 18, 19, 24, 27	190

Table 6 Comparison of evaluation indicators before and after improvement

Stage	Production balance indicators				
	CT	SI	d	k	Objective function
Pre-improvement	221	36.53	0.14	0.25	0.49
Improved	195	7.87	0.03	0.10	0.53

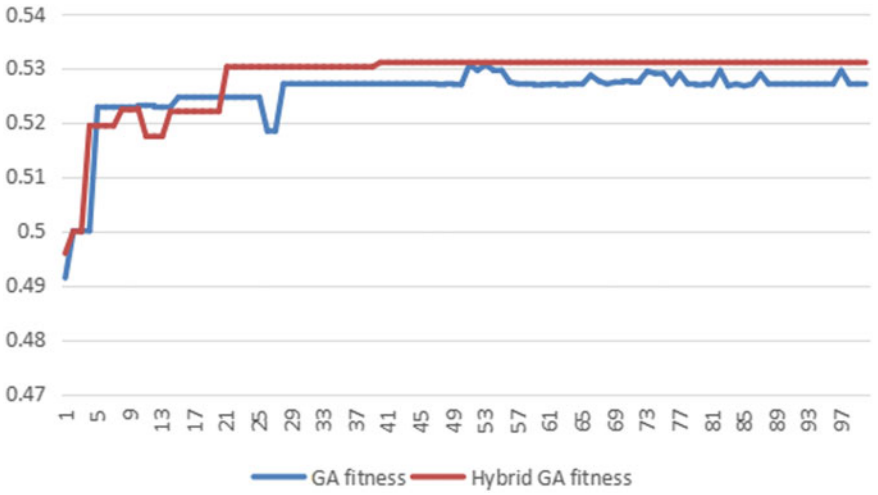


Fig. 8 Iteration curve

The equilibrium loss coefficient changed from 14.4 to 3% and the improved equilibrium loss coefficient is less than 10%, proving that the assembly line is currently in "excellent" condition, an improvement of 11.4%.

2. *Evaluation of the convergence effect of the algorithm*

The balancing optimization of the mixed-flow assembly line for the instrument assembly is performed separately using a traditional genetic algorithm and a hybrid genetic algorithm combined with simulated annealing, with the number of iterations both set to 100, thus allowing the cycle time of the assembly line to be reduced and the second type of balancing problem to be optimized.

The iterative curves for the application of the two algorithms for optimization are shown in Fig. 8, and the application of the mixing and genetic algorithm for optimization of the mixed-flow assembly line in the population tends to equilibrate around 20 generations of population iterations, while the application of traditional genetic algorithms tends to equilibrate around 30 generations of population iterations. Therefore, for solving the Type II equilibrium problem of the mixed-flow assembly line, the hybrid genetic algorithm combined with the simulated annealing algorithm converges faster. Optimized results can be obtained more quickly.

6 Conclusion

In this paper, a mixed-flow assembly line equilibrium model incorporating a hybrid genetic algorithm for simulated annealing was developed by MATLAB, and a company's equilibrium optimization for instrument assembly lines. For the disadvantages of the genetic algorithm in the optimization process, which is prone to fall into the local optimum solution and precociousness, the article focuses on the second type of balancing for the mixed-flow assembly line issue. A genetic algorithm was designed, taking into account three evaluation metrics: the smoothing index, the equilibrium loss coefficient and the imbalance coefficient. The simulated annealing algorithm is combined with a genetic algorithm to speed up convergence to obtain a global optimal solution, which is later verified by example analysis feasibility of the method.

Finally, the example is verified by using the instrument mixed-flow assembly line of Company L. According to the objective function composed of the evaluation index smoothness index, balance loss coefficient and unbalance coefficient, the comparison before and after improvement is made. The value of cycle time on the assembly line improved from 221 to 195 s when the objective function consisting of the balance coefficient was compared before and after improvement. The smoothness index changed from 36.53 to 7.87, a decrease of 28.66. The balance loss coefficient changed from 14.4 to 3%, and the value of cycle time for the assembly line improved from 221 to 195. Effectively solved the balance of the company's instrument assembly line.

Meanwhile, the optimization results of the traditional genetic algorithm and the hybrid genetic algorithm are compared to form an iterative graph of the optimization process of the two algorithms, with the hybrid genetic algorithm converging faster, which proves that the hybrid genetic algorithm has certain advantages for solving the balance problem of the mixed-flow assembly line, and also proves that the assembly line balance model established in this paper can solve such production line problems faced by the manufacturing enterprise, improve the production efficiency, and improve the economic efficiency and market competitiveness of the enterprise.

It should be noted that the model in this paper is based on the conditions of the assembly line determination and does not take into account uncertainties such as equipment failure, the employee proficiency, order insertion/change order production, product changeover time, etc. In order to be able to more effectively achieve precision operations on mixed assembly lines, combining the deterministic and indeterministic conditions will be the next step in this paper work focus.

Acknowledgements This work was partially supported by Beijing Logistics Informatics Research Base and Intelligent Emergency Logistics Linkage System of Public Emergencies in Beijing under Grant 18JDGLB019.

References

1. X. Li, S. Dong, Study of hybrid genetic algorithm for the balance problem of mixed flow assembly line. J. Univ. Sci. Technol. Beijing **34**, 952–958 (2012)
2. K. Agpak, H. Gokcen, Assembly line balancing: two cases of resource constraints. Prod. Econ. **96**(1), 129–140 (2007)
3. M. Held, R.M. Karp, R. Shareshian, Assembly- line balancing-dynamic programming with precedence constraints. Oper. Res. **11**(3), 442–459 (1963)
4. K. Li, Study on the optimization of the balance ratio of production line in H based on 0-1 integer planning. Intern. Combust. Engine Parts **24**, 143–144 (2018)
5. Manhui He, Kai Zheng, Research on the balance improvement of PC machine assembly production line using MOD method. Modern Manuf. Eng. **7**, 51–55 (2017)
6. K. Zhang, Research on the optimal design of furniture production line based on genetic algorithm. J. Chifeng College (Nat. Sci. Ed.) **35**(6), 111–113 (2019)
7. Xujing Zhang, Lichuan Wang, Yan Chen, Genetic algorithm-based balance optimization of garment sewing production line. J. Text. **41**(2), 125–129 (2020)
8. Jingfang Fang, Xu Yankai, Production line balance optimization based on iterative search and genetic algorithm. Comput. Appl. Softw. **34**(8), 276–280 (2017)
9. Squire Jing, Jiang Enhance, MaoGen Ge, Research on the balance problem of hybrid assembly line based on improved genetic algorithm. J. Hefei Univ. Technol. (Nat. Sci.) **33**(7), 1006–1009 (2010)
10. B.Q. Yang, M.S. Tong, Mathematical modeling of automobile assembly line balancing. Electr. Drive Autom. **26**(1), 60 (2004)
11. Bengang Wang, Batch and sequence integration optimization study of mixed flow production line based on GASA with buffer constraints. Modern Manuf. Eng. **2**, 73–77 (2016)
12. J. Tang, Y. Yang, Y. Qi, A hybrid algorithm for urban transit schedule optimization. Physica, A. Stat. Mech. Appl. **5122**, 745–755 (2018)
13. L. Wang, *Intelligent Optimization Algorithm and Its Application* (Tsinghua University Press, Beijing, 2001)
14. Dan Chang, Lizhu Cui, Zhao Huang, A cellular-automaton agent-hybrid model for emergency evacuation of people in public places. IEEE Access **8**, 79541–79551 (2020)
15. D. Chang, R. Fan, Z.T. Sun, A deep belief network and case reasoning based decision model for emergency rescue. Int. J. Comput. Commun. Control **15**(3), 3836 (2020)
16. M.S. Yang, L. Ba, H.Y. Zheng, Y. Liu, X.F. Wang, J.Z. He, Y. Li, An integrated system for scheduling of processing and assembly operations with fuzzy operation time and fuzzy delivery time. Adv. Prod. Eng. Manage. **14**(3), 367–378 (2019)
17. C. Jiang, J.T. Xi, Dynamic scheduling in the engineer-to-order (ETO) assembly process by the combined immune algorithm and simulated annealing method. Adv. Prod. Eng. Manage. **14**(3), 271–283 (2019)
18. S.L. Yang, Z.G. Xu, J.Y. Wang, Modelling and production configuration optimization for an assembly shop. Int. J. Simul. Model. **18**(2), 366–377 (2019)
19. M.S. Yang, L. Ba, Y. Liu, H.Y. Zheng, J.T. Yan, X.Q. Gao, J.M. Xiao, An improved genetic simulated annealing algorithm for stochastic two-sided assembly line balancing problem. Int. J. Simul. Model. **18**(1), 175–186 (2019)
20. I. Vrecko, J. Kovac, B. Rupnik, B. Gajsek, Using Queuing simulation model in production process innovations. Int. J. Simul. Model. **18**(1), 47–58 (2019)
21. H. Wu, A. Tsai, H. Wu, A hybrid multi-criteria decision analysis approach for environmental performance evaluation: an example of the TFT-LCD manufacturers in Taiwan. Environ. Eng. Manage. J. **18**, 597–616 (2019)
22. E. Comăniță, P. Cozma, I. Simion, M. Roșca, M. Gavrilescu, Evaluation of eco-efficiency by multicriteria decision analysis. Case study of eco-innovated and eco-designed products from recyclable waste. Environ. Eng. Manage. J. **17**, 1791–1804 (2018)

Informatization Development Mode of Internal Control in Colleges and Universities Based on Principal-Agent

Yuzhen Cheng

Abstract Informatization of internal control is a critical part of internal control construction. It is a great challenge faced by administrative institutions and colleges and universities in terms of how to realize the informatization of internal control. Informatization of internal control in colleges and universities requires systematic work, in which the choice of system development mode is the key. This paper, based on the informatization construction of internal control in administrative institutions, innovatively introduces the principal-agent theory. This paper discusses the selection of the system development mode under the background of informationization of internal control in colleges and universities. This paper compares the joint deterministic equivalent surplus of two different task allocationed modes, and uses the case of X university to corroborate accordingly. The results show that it is a better choice to completely allocate the task to an agent and it is also an effective incentive mode in Colleges and Universities. This paper's research on system development mode can provide theoretical support for informatization of internal control in administrative institutions.

Keywords Internal control · Principal-agent · Informatization · Development mode

1 Introduction

In order to improve the management standards of administrative institutions, the Ministry of Finance began to implement the "Internal Control Standard for Administrative Institutions (Trial Implementation)" Accounting [2012] No. 21) (hereinafter referred to as the "Internal Control Standard") in 2014. The "Internal Control Standard" aimed to improve risk control of administrative institutions and to build a better

Y. Cheng (✉)
Office of Financial Affairs, Beijing Jiaotong University, Beijing, China
e-mail: yzcheng@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_65

informatization of internal control management system. The "Internal Control Standards" stated that "Administrative institutions should make full use of modern scientific techniques and tools to strengthen internal control. Administrative institutions should implement centralized management of information system construction, and embed the economic activity and its internal control processes into institutions' information system, so as to reduce or eliminate human manipulation factors and protect information security [1]". The Ministry of Education formulated "the Guideline for Internal Control of Economic Activities of Colleges and Universities Directly Under the Ministry of Education (Trial Implementation)" (hereinafter referred to as the "Internal Control Guideline") in 2016. The "Internal Control Guideline" proposed management methods and related regulations of internal control in colleges and universities, to improve the construction of informatization of internal control system, and pay attention to internal control from the development strategy. The guideline also put forward the requirements on "Information Management Construction of Internal Control of Economic Activities [2]". In October 2016, the Ministry of Finance issued the "Outline of Thirteenth Five-Year Plan for Accounting Reform and Development" [3]. The document proposed the accounting informatization requirements for public institutions and enterprises in China to avoid internal and external risks and ensure stability through continuous improvement of the informatization of internal control system. In recent years, the state has promulgated a number of laws and regulations on internal control of administrative institutions, which shows that the state pays high attention to internal control and its informatization.

With the comprehensive development of "Internet plus" and informatization, administrative institutions, especially colleges and universities, have established various informatization management systems. However, the construction and management of informatization of internal control is not yet mature. How to implement informatization of internal control is a major problem generally faced by administrative institutions and even colleges and universities. Informatization of internal control in colleges and universities is a systematic work, and it needs to be carried out in accordance with the standardized implementation step under the guidance of scientific methods. And the selection of system development mode is the key process, which is constrained by conditions such as cost, applicability, practicability, scalability, ease of maintenance, and security.

The "Internal Control Guideline" mentions that "Colleges and universities can adopt the modes of departmental self-development, joint development of multiple departments, direct outsourcing of commercial software, or commissioning of customized development by external companies, to develop economic activity information systems. Colleges and universities should fully evaluate the development and maintenance risks of selected modes, and prepare risk treatment plans [2]". Self-development and joint development of multiple departments belong to the "user development mode", and direct outsourcing of commercial software and commissioning of custom development by external entities belong to the "delegation development mode". "User development mode" is suitable for colleges and universities with a strong IT team, which saves costs and facilitates later system maintenance. "Delegation development mode" has low requirements for the management level and personnel quality of colleges and universities internal control, but it has certain costs and requires the participation of development company personnel with later system maintenance. The paper introduces the principal-agent theory and takes internal control as a control mechanism, including the owner's monitoring over the operator and the operator's monitoring on the business process. Thereby it protects the principal's interests and improves the principal-agent relationship. From the perspective of informatization of internal control, informatization centralized management department of internal control serves as the principal, and the system development company serves as the agent. Their behaviors can be monitored by the management of internal control. Based on the principal-agent theory, this paper discusses the introduction of the principal-agent model in system development mode selection of informatization of internal control.

2 Literature Review

The research on informatization of internal control in foreign countries is relatively complete. Marshall [4] put forward an innovative point of view, that the construction of information system already included a part of internal control. Because the collection and storage of data in some degree can analyze the existence of problems in the company's internal control management. At the same time, he also proposed that informatization of internal control can improve the effectiveness of internal management. But the information system affected internal control in other aspects, and it needed in-depth empirical study to determine this. Yang and Koo [5] believed that perfecting the internal control mechanism needs to gradually build IT systems. However, the adoption of new IT systems often leads to internal control defects and operational risks. The authors proposed to use IT systems to build an internal control adaptation model (ICAM). The research results provide a reference for building a new internal control mechanism and theoretical basis for further research. Popescu et al. [6] believed that the internal management and control system of public entities includes three objectives: the effectiveness and efficiency of operations, the reliability of internal and external information, and the compliance with relevant laws, regulations and internal policies. The implementation and development of the internal control system can be promoted by developing and updating the internal control system. The domestic researches on informatization of internal control started late, mainly focusing on the research of information management form and legal level, and the development of theory is promoted through practice. However, some researchers have begun to study the framework of building internal control information systems. Huayang and Tong [7] believed that the internal control information system is the only way for continuous improvement of internal control management. The essence of internal control is to review and control processes and financial data. However, there are also many problems in the internal control information system, such as the cost of informatization, computer talent reserves, system maintenance, and data security. Finally, it proposed that continuous innovations of technologies, such as cloud computing, will continuously improve the ability of the internal control information system. Guocheng and Lili [8] analyzed the causes of the risk of Meso Information System and believed that data mining algorithms should be introduced to make the fuzzy control system clear, accurate and measurable. The model designed the internal control management system in line with the information system of China's mid-level economy.

Most domestic institutions often choose the development method depending on the empirical judgment of the leaders and informatization personnel, which lacks a certain scientific nature. Juan [9], from the perspective of quantitative analysis, through the AHP method and with the help of Expert Choice software, respectively proposed the most suitable software development method for six different types of enterprises. Wei [10] proposed that logistics companies must choose their own development methods according to their own business, technical personnel and capital status from a qualitative point of view. In this way, it could improve the efficiency of logistics companies and enhance their competitiveness. In terms of the principal-agent model, Holmstrom and Milgrom [11] analyzed the principal-agent model, described of the linear principal-agent model, allocated incentives for effort and attention, and limited on outside activities. Zhen [12] established a win-win cooperative principal-agent model for projects of Public-private Partnerships, and introduced the Cobb Douglas production function as the output function in the model. She analyzed the balanced benefit distribution ratio obtained by the model and the level of balanced effort enterprises achieved. Hu [13] introduced the organizational structure into the principal-agent model and constructed a dual agent model with the organizational structure as an endogenous variable. Kjell [14] mentioned that principal–agent theory and game theory are applied to the precautionary principle (PP) to open up a new research agenda. Bi et al. [15] believed that the principal–agent theory is used to control the schedule risk of ITO projects. This paper built a two-level mathematical model to describe the decision process of the client and vendors, and designed a genetic algorithm (GA) to solve the proposed model.

3 Model of Informatization Development of Internal Control

Based on the linear model of Holmstrom and Milgrom [11], this paper assumes that informatization centralized management department of internal control serves as the principal and the system development company serves as the agent. As the representative of the interests of colleges and universities, the principal aims to maximize the total social surplus of state-owned assets and user satisfaction on the basis of completing informatization of internal control tasks through certain incentives. We assume that the agent is a "rational economic man" with its own interest demands, and has separate departmental interests. Part of the department's profits come from the compensation (that is, the remuneration or wages) paid by the state, and may also obtain more profits through rent-seeking behavior. As the agent, the system development company costs to exercise internal control, and has a vector of efforts $y = (y_1, y_2, \ldots, y_i)$, which indicates how much efforts it puts into tasks.

We assume that informatization of internal control centralized management department wants the agent to carry out two tasks,$m(m = 1, 2)$, and the utility obtained through work is $x = y_1 + y_2 + \varepsilon$. $ym(y_m \geq 0)$ is the efforts that system development company n has devoted to the task m, which informatization centralized management department of internal control cannot directly observe. The system development company has common information and the ε is normally distributed with mean zero and variance $\sigma^2 > 0$. Since the efforts of the system development company cannot be observed by the principal, we assume that the profits gained from the work is the only information that can be reflected in the principal-agent contract. Informatization centralized management department of internal control mainly implements incentives by paying commissions and the choices of task allocation to the system development company. At the same time we assume that informatization centralized management department of internal control also has the ability to complete the work, but cannot complete the task alone. Therefore, informatization centralized management department of internal control can consider two different task allocationed modes: (1) Local Delegation, which indicates that informatization of internal control centralized management department assigns one of the tasks to the system development company and undertakes another task by itself. (2) Complete Delegation, which indicates that informatization of internal control centralized management department delegates the tasks to different system development companies.

$b(x)$ denotes the commission paid by the system development company in the Local Delegation. $b_n(x)$ denotes the commission paid by the system development company in the Complete Delegation. In order to facilitate the analysis, we assume that the cost function of informatization centralized department of internal control to complete the task is quadratic $F(y_1, y_2) = \frac{1}{2}gy_1^2 + \frac{1}{2}gy_2^2 + \delta g y_1 y_2$. The degree of difficulty of the task is $g > 0$ and the parameter $\delta \in [0, 1]$ indicates the degree of cost of substitution between different tasks. If $\delta = 0$, then the two tasks are independent of each other; If $\delta > 0$, then increasing the effort on task 1 would increase the marginal costs of investing effort on task 2; If $\delta = 1$, then the two tasks could be completely substitutes, and the cost would only depend on the total level of effort $(g_1 + g_2)$. We assume that the cost function of the two tasks is of the same form $F_1(y) = F_2(y) = \frac{1}{2}gy^2$. Due to the symmetry, it is only necessary to consider the case where the system development company accepts task 1 in the Local Delegation. $A \in [p, q]$ denotes the assignment mode of the tasks. $A = p$ is Local Delegation, and $A = q$ is Complete Delegation.

3.1 Delegation Mode p: Local Delegation

Informatization centralized management department of internal control assigns task 1 to the system development company, and undertakes another task 2 by itself. The commission paid to the system development company is $b(x)$. The payment of commission is linear of the form $b(x) = \alpha x + \gamma$. The cost of completing the assigned task 1 by the system development company is $F_1(y_1) = \frac{1}{2}gy_1^2$, and the cost of completing the task 2 of informatization of internal control centralized management department is $F_2(y_2) = \frac{1}{2}gy_2^2$. We assume that informatization centralized management department of internal control is risk neutral, but the system development company is a risk avoider. The system development company's preference can be expressed as an exponential utility function $-\exp\{-\gamma I\}$. $\gamma > 0$ is the absolute risk aversion coefficient. I is net income (payments received minus costs). At the same time, we assume that the system development on the list has equal absolute risk aversion coefficient and net work efficiency. The acceptable retention payment to the system development company is 0. The informatization centralized management department of internal control maximizes the total certainty equivalent subject to the incentive compatibility constraints by selecting the task allocation parameters α or (α_1, α_2). The joint deterministic equivalent surplus is:

$$y_1 + y_2 - F_1(y_1) - F_2(y_2) - \frac{1}{2}\gamma\sigma^2\alpha^2 \tag{1}$$

Incentive compatibility constraints are:

$$\alpha - F_1'(y_1) = 0 \tag{2}$$

$$(1 - \alpha) - F_2'(y_2) = 0 \tag{3}$$

Next step is to solve the problem of the task's optimal commission ratio. In the incentive compatibility constraint, the maximal joint deterministic equivalent surplus obtains the optimal delegation ratio. The model is:

$$\begin{aligned} &\max_{\alpha} \quad y_1 + y_2 - \frac{1}{2}gy_1^2 - \frac{1}{2}gy_2^2 - \frac{1}{2}\gamma\sigma^2\alpha^2 \\ &\text{s.t.} \quad \alpha - gy_1 = 0 \\ &(1 - \alpha) - gy_2 = 0 \end{aligned}$$

3.2 Delegation Mode q: Complete Delegation

Informatization centralized management department of internal control assigns task 1 to system development company 1, and task 2 to system development company 2. The commission paid by informatization centralized management department of internal control to the system development company 1 is $b_1(x)$, and the commission payment is linear of the form $b_1(x) = \alpha_1 x + \gamma$. The commission paid by informatization centralized management department of internal control to the system development company 2 is $b_2(x)$, and the commission payment adopts the form of a linear function $b_2(x) = \alpha_2 x + \gamma$. The cost incurred by system development company 1 to complete assigned task 1 is $F_1(y_1) = \frac{1}{2} g y_1^2$, and by system development company 2 to complete task 2 is $F_2(y_2) = \frac{1}{2} g y_2^2$. Other assumptions are the same as Local Delegation. The joint deterministic equivalent surplus is:

$$y_1 + y_2 - F_1(y_1) - F_2(y_2) - \frac{1}{2}\gamma\sigma^2\alpha_1^2 - \frac{1}{2}\gamma\sigma^2\alpha_2^2 \tag{4}$$

Incentive compatibility constraints are:

$$\alpha_1 - F_1'(y_1) = 0 \tag{5}$$

$$\alpha_2 - F_2'(y_2) = 0 \tag{6}$$

Next step is to solve the problem of the task's optimal commission ratio. In the incentive compatibility constraint, the maximal joint deterministic equivalent surplus obtains the optimal delegation ratio. The model is:

$$\begin{aligned}
&\max_{\alpha_1,\alpha_2} \quad y_1 + y_2 - \frac{1}{2} g y_1^2 - \frac{1}{2} g y_2^2 - \frac{1}{2}\gamma\sigma^2\alpha_1^2 - \frac{1}{2}\gamma\sigma^2\alpha_2^2 \\
&\text{s.t. } \alpha_1 - g y_1 = 0 \\
&\alpha_2 - g y_2 = 0
\end{aligned}$$

4 Analysis and Solution of Model

4.1 Delegation Mode p: Local Delegation

It is solved by constraints 2 and 3: $y_1 = \frac{\alpha}{g}$, $y_2 = \frac{1-\alpha}{g}$. They are substituted into the objective function to get $\max_\alpha \frac{1}{2g}\left[1 + 2\alpha - \alpha^2\left(2 + \gamma\sigma^2 g\right)\right]$.

Optimal task allocation parameter is:

$$\alpha_p^* = \frac{1}{2+\gamma\sigma^2 g}$$

Substituting α_p^* into the constraints 2 and 3 we can get

$$y_1^* = \frac{1}{g\left(2+\gamma\sigma^2 g\right)}, \quad y_2^* = \frac{1}{g}\left(1 - \frac{1}{2+\gamma\sigma^2 g}\right)$$

Substituting α_p^* into the objective function 1, we can get the joint deterministic equivalent surplus

$$\begin{aligned} &\frac{1}{2g}\left[1 + 2\alpha_p^* - \alpha_p^{*2}\left(2+\gamma\sigma^2 g\right)\right] \\ &= \frac{1}{2g}\left[1 + 2 * \frac{1}{2+\gamma\sigma^2 g} - \left(\frac{1}{2+\gamma\sigma^2 g}\right)^2 \left(2+\gamma\sigma^2 g\right)\right] \\ &= \frac{1}{2g}\left(1 + \frac{1}{2+\gamma\sigma^2 g}\right) \end{aligned}$$

Theorem 1 can be obtained from the above analysis.

Theorem 1 *In the local delegation mode, the optimal task allocation parameter of the commission paid by informatization centralized management department of internal control to the system development company is* $\alpha_p^* = \frac{1}{2+\gamma\sigma^2 g}$*, and the joint deterministic equivalent surplus is* $\frac{1}{2g}\left(1 + \frac{1}{2+\gamma\sigma^2 g}\right)$.

The following analysis is sensitivity analysis of the optimal task allocation parameter α_p^* with respect to the parameter g,γ and σ^2 in the commission mode p.

Property 1 *The optimal task allocation parameter* α_p^* *in the delegation mode p decreases with the increase of* g.

The proof is as follows: $\frac{\partial \alpha_p^*}{\partial g} = -\frac{\gamma\sigma^2}{\left(2+\gamma\sigma^2 g\right)^2} < 0$. Therefore, the optimal task allocation parameter α_p^* in the delegation mode p decreases with the increase of g.

As the parameter $g > 0$, theorem 1 shows that when g approaches infinitely lower bound 0, the limit of the optimal task allocation parameter α_p^* is $\frac{1}{2}$. And when g approaches infinity, the limit of the optimal task allocation parameter α_p^* is 0, where the joint effect cannot be measured.

In summary, according to Property 1, we can get the optimal task allocation parameter $\alpha_p^* \in \left(0, \frac{1}{2}\right)$ when $g \in (0, +\infty)$ in the delegation mode p.

From the constraints we know that $y_1 = \frac{\alpha}{g}$, $y_2 = \frac{1-\alpha}{g}$, then $\frac{\partial t_1}{\partial \alpha} = \frac{1}{g}$, $\frac{\partial t_2}{\partial (1-\alpha)} = \frac{1}{g}$. The parameter g denotes the degree of difficulty of completing a task, then $\frac{1}{g}$ represents the response to the effort of completing the task with incentives. According to Property 1, the optimal task allocation parameter α_p^* in the delegation mode p decreases with the increase of g, which means the optimal task allocation parameter α_p^* increases with the increase of $\frac{1}{g}$. If the system development company has a strong response to the incentives, it will receive a stronger incentive, and this incentive response is controllable. Therefore informatization centralized management department of internal control can achieve more effective incentives by reducing the task difficulty level and improving the working conditions of the system development company.

Property 2 *The optimal task allocation parameter α_p^* in the delegation mode p decreases with the increase of the absolute risk aversion coefficient γ.*

The proof is as follows:

$$\frac{\partial \alpha_p^*}{\partial \gamma} = -\frac{\sigma^2 c}{\left(2 + \gamma \sigma^2 g\right)^2} < 0.$$

Therefore, the optimal task allocation parameter α_p^* in the delegation mode p decreases with the increase of the parameter γ.

The coefficient $\gamma > 0$, so Theorem 1 shows that when γ approaches infinitely lower bound 0, the limit of the optimal task allocation parameter α_p^* is $\frac{1}{2}$. When γ approaches infinity, the limit of the optimal task allocation parameter α_p^* is 0, therefore the joint effect cannot be measured.

In summary, according to Property 2, we can get the optimal task allocation parameter $\alpha_p^* \in \left(0, \frac{1}{2}\right)$ when $\gamma \in (0, +\infty)$ in the delegation mode p.

Property 3 *The optimal task allocation parameter α_p^* in the delegation mode p decreases with the increase of the variance σ^2.*

The proof is as follows:

$$\frac{\partial \alpha_p^*}{\partial \sigma^2} = -\frac{\gamma g}{\left(2 + \gamma \sigma^2 g\right)^2} < 0.$$

Therefore, the optimal task allocation parameter α_p^* in the delegation mode p decreases with respect to the variance σ^2.

Since the variance σ^2 can measure the difficulty of the combined effect, and $\gamma > 0$, Theorem 1 shows that when σ^2 approaches infinitely lower bound 0, the limit of the optimal task allocation parameter α_p^* is $\frac{1}{2}$. When σ^2 approaches infinity, the limit of the optimal task allocation parameter α_p^* is 0, therefore the joint effect cannot be measured.

In summary, according to Property 3, it gets that the optimal task allocation parameter $\alpha_p^* \in \left(0, \frac{1}{2}\right)$, when $\sigma^2 \in (0, +\infty)$ in the delegation mode p.

4.2 Delegation Mode q: Complete Delegation

It is solved by constraints 5 and 6: $y_1 = \frac{\alpha_1}{g}$, $y_2 = \frac{\alpha_2}{g}$. They are substituted into the objective function to get

$$\max_{\alpha_1,\alpha_2} \frac{1}{2g}\left[2\alpha_1 - \alpha_1^2\left(1+\gamma\sigma^2 g\right) + 2\alpha_2 - \alpha_2^2\left(1+\gamma\sigma^2 g\right)\right].$$

Optimal task allocation parameter is:

$$\alpha_{q1}^* = \frac{1}{1+\gamma\sigma^2 g}, \alpha_{q2}^* = \frac{1}{1+\gamma\sigma^2 g}.$$

Substituting α_{q1}^*, α_{q2}^* into the constraints 5 and 6, we can get

$$y_1^* = y_2^* = \frac{1}{g\left(1+\gamma\sigma^2 g\right)}.$$

Substituting α_{q1}^*, α_{q2}^* into the objective function 4 gets the joint deterministic equivalent surplus:

$$\begin{aligned}
&\frac{1}{2g}\left[2\alpha_{q1}^* - \alpha_{q1}^{*2}\left(1+\gamma\sigma^2 g\right) + 2\alpha_{q2}^* - \alpha_{q2}^{*2}\left(1+\gamma\sigma^2 g\right)\right] \\
&= \frac{1}{2g}\left[\begin{array}{l} 2 * \frac{1}{1+\gamma\sigma^2 g} - \left(\frac{1}{1+\gamma\sigma^2 g}\right)^2\left(1+\gamma\sigma^2 g\right) + 2 * \frac{1}{1+\gamma\sigma^2 g} \\ -\left(\frac{1}{1+\gamma\sigma^2 g}\right)^2\left(1+\gamma\sigma^2 g\right) \end{array}\right] \\
&= \frac{1}{g\left(1+\gamma\sigma^2 g\right)}
\end{aligned}$$

Theorem 2 can be obtained from the above analysis.

Theorem 2 *In the complete delegation mode, the optimal task allocation parameter of the commission paid by informatization centralized management department of internal control to the system development company is* $\alpha_{q1}^* = \alpha_{q2}^* = \frac{1}{1+\gamma\sigma^2 g}$, *and the joint deterministic equivalent surplus is* $\frac{1}{g\left(1+\gamma\sigma^2 g\right)}$.

The following analyze is sensitivity analysis of the optimal task allocation parameter α_{q1}^*, α_{q2}^* with respect to the parameter g, γ and σ^2 in the commission mode p.

Property 4 *The optimal task allocation parameter* $\alpha_{q1}^* = \alpha_{q2}^*$ *in the delegation mode p decreases with increase of the difficulty degree* g.

The proof is as follows:

$$\frac{\partial \alpha_{q1}^{*}}{\partial g} = \frac{\partial \alpha_{q2}^{*}}{\partial g} = -\frac{\gamma \sigma^{2}}{\left(1 + \gamma \sigma^{2} g\right)^{2}} < 0.$$

Therefore, the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ in the delegation mode p decreases with respect to the parameter g.

As the parameter $g > 0$, Theorem 2 shows that when g approaches infinitely lower bound 0, the limit of the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ is 1. And when g approaches infinity, the limit of the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ is 0, when the joint effect cannot be measured.

In summary, according to Property 4, we can get the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*} \in (0, 1)$, when $g \in (0, +\infty)$ in the delegation mode p.

From the constraints we know that $y_1 = \frac{\alpha}{g}$, $y_2 = \frac{1-\alpha}{g}$, then $\frac{\partial t_1}{\partial \alpha_1} = \frac{\partial t_2}{\partial \alpha_2} = \frac{1}{g}$. The parameter g denotes the degree of difficulty of completing a task, then $\frac{1}{g}$ represents the response to the effort for completing the task with incentives. According to Property 4, the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ in the delegation mode p decreases with the increase of g, which means the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ increases with the increase of $\frac{1}{g}$. If the system development company has a strong response to the incentive, it will receive a stronger incentives, and this incentive response is controllable. Therefore informatization centralized management department of internal control can achieve more effective incentives by reducing the task difficulty level and improving the working conditions of the system development company.

Property 5 *The optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ in the delegation mode p decreases with the increase of the absolute risk aversion coefficient γ.*

The proof is as follows:

$$\frac{\partial \alpha_{q1}^{*}}{\partial \gamma} = \frac{\partial \alpha_{q2}^{*}}{\partial \gamma} = -\frac{\sigma^{2} g}{\left(1 + \gamma \sigma^{2} g\right)^{2}} < 0.$$

Therefore, the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ in the delegation mode p decreases with the increase of the parameter γ.

The coefficient $\gamma > 0$, so Theorem 2 shows that when γ approaches infinitely lower bound 0, the limit of the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ is 1. When γ approaches infinity, the limit of the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ is 0, therefore the joint effect cannot be measured.

In summary, according to Property 5, we can get the optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*} \in (0, 1)$ when $\gamma \in (0, +\infty)$ in the delegation mode p.

Property 6 *The optimal task allocation parameter $\alpha_{q1}^{*} = \alpha_{q2}^{*}$ in the delegation mode p decreases with the increase of the variance σ^{2}.*

Table 1 Comparison table of parameters in delegation mode

Delegation mode	Task assignment parameters	Joint deterministic equivalent surplus
Local delegation	$\alpha_p^* = \frac{1}{2+\gamma\sigma^2 g}$	$\frac{1}{2g}\left(1 + \frac{1}{2+\gamma\sigma^2 g}\right)$
Complete delegation	$\alpha_{q1}^* = \alpha_{q2}^* = \frac{1}{1+\gamma\sigma^2 g}$	$\frac{1}{g(1+\gamma\sigma^2 g)}$

The proof is as follows:

$$\frac{\partial\alpha_{q1}^*}{\partial\sigma^2} = \frac{\partial\alpha_{q2}^*}{\partial\sigma^2} = -\frac{\gamma g}{\left(1+\gamma\sigma^2 g\right)^2} < 0.$$

Therefore, the optimal task allocation parameter $\alpha_{q1}^* = \alpha_{q2}^*$ in the delegation mode p decreases with the increase of the variance σ^2.

Because the variance σ^2 can measure the difficulty of the combined effect, and $\sigma^2 > 0$, therefore Theorem 2 shows that when σ^2 approaches infinitely lower bound 0, the limit of the optimal task allocation parameter $\alpha_{q1}^* = \alpha_{q2}^*$ is 1. When σ^2 approaches infinity, the limit of the optimal task allocation parameter $\alpha_{q1}^* = \alpha_{q2}^*$ is 0, therefore the joint effect cannot be measured.

In summary, according to Property 6, we can get the optimal task allocation parameter $\alpha_{q1}^* = \alpha_{q2}^* \in (0, 1)$ when $\sigma^2 \in (0, +\infty)$ in the delegation mode p.

4.3 Conclusion and Summary

In summary, according to Theorems 1 and 2, Table 1 can be obtained.

This paper further analyzes and compares the optimal task allocation parameters of the local delegation mode and the complete delegation mode.

$$\begin{aligned} \alpha_p^* - \alpha_{q1}^* &= \alpha_p^* - \alpha_{q2}^* \\ &= \frac{1}{2+\gamma\sigma^2 g} - \frac{1}{1+\gamma\sigma^2 g} \\ &= \frac{-1}{\left(2+\gamma\sigma^2 g\right)\left(1+\gamma\sigma^2 g\right)} \\ &< 0 \end{aligned}$$

Therefore, the optimal task allocation parameters of the local delegation mode are smaller than those of the full delegation mode. However, it is necessary to simultaneously provide incentives to informatization centralized management department of internal control and system development company. In fact, increasing the payment to the system development company will correspondingly reduce the enthusiasm of informatization centralized management department of internal control engaged in the task. The limit of this distribution mode is completely delegationed.

By comparing the joint surplus of allocation modes of two tasks, the following conclusions can be drawn. For any $\delta \in [0, 1]$, when the values of γ and σ^2 are small enough, the joint deterministic equivalent surplus of the local delegation mode approaches $\frac{3}{4g}$, and that of the complete delegation mode approaches $\frac{1}{g}$. Therefor the complete delegation mode is superior for local delegation mode.

The conclusion shows that, compared with the department to completing the task by itself, it is a better option to delegate the task to the agent. It is also an effective incentive method to delegate the task to the system development company. And, by all means, task allocation method of the complete delegation will incur corresponding costs, and the system development company takes higher risks due to full responsibility.

5 Case Study: X University's Development Mode

Based on the above analysis, the complete delegation is a relatively optimal task allocation model. This delegation not only provides theoretical explanation and basis for the development model of informationization of internal control, but also can be corroborated from the case of colleges and universities.

This paper takes X university as an example of administrative institutions. The internal control management system of X university is relatively mature, and X university combines it with the information system to carry out informationization construction of internal control. Under the requirements of informatization of internal control, on the basis of the existing information system, X university first improves the six business management module system. Then it establishes informatization system of internal control. And at the same time, through the support of the integrated database of big data, processes data uniformly. X university relies on informatization system of internal control to process the integrated data logically. The result is finally output to the corresponding internal control nodes and timely feeds back to the corresponding staff of supervision and evaluation of internal control. It can effectively monitor and evaluate each risk node in real time, so as to solve the corresponding risk problems faster. In the first phase of construction, X university built a contract management system and a data exchange platform of internal control.

Before the implementation of informatization construction of internal control, asset management system of X university was a system development task undertaken by informatization office itself. The financial platform was allocated to the system development company by the finance office. This development model adopted the local delegation mode. In actual use, it is found that there is an information island between the financial platform and the asset management system. Due to system compatibility issues between the two systems, data cannot be shared. In addition, the asset management system was limited by the input of personnel in the informationization office of X university, so the long development cycle of the system and the untimely maintenance of the system affected the normal use of the system.

In the first phase of informatization construction of internal control, the informatization centralized management department of internal control in A university adopted the complete delegation mode, that is, the two development tasks were completely allocated to two system development companies. System development company 1 is committed to the research and development of related products in the education industry, and has developed a contract management system for X university. The system is extensible and easy to maintain, and it is developed according to the actuality of contract management of X university. It promotes process and refined management, and improves the contract management standards of X university. System development company 2 has a full-process and big data independent platform product, which has data collection, processing and analysis capabilities. It has developed a data exchange platform of internal control for X university, which has broken the information island and realized the mutual reception and transmission of data between systems in various departments. This has improved the utilization rate of data and the work efficiency of various departments and ensured the accuracy of data acquisition. By the deployment of the contract management system and data exchange platform of internal control in X university, the problems of development cycle and scalability were solved, and various departments responded well.

During the project acceptance, the review experts of the project evaluated the influencing factors of the two different modes. The evaluation results are shown in Table 2. In the complete delegation mode, the system development cycle is fast, the R&D team is expanded, and the system is more expandable and practical. But the relative development cost is more. And for colleges and universities, the development cycle, system scalability, and practicality are more important than the cost. In summary, the informationization construction of internal control in X university has proved in practice that, compared with the department to complete the task by itself, it is a relatively optimal task allocation model to completely allocate the task to the system development company. Therefore, the complete delegation mode is applicable to colleges and universities as administrative institutions.

Table 2 Comparison table of influencing factors in development mode

Delegation mode	Development cycle (years)	R&D staff (people)	Cost (ten thousand yuan)	System scalability (10 points total)	System suitability (10 points total)
Local delegation	2	4	30	6	7
Complete delegation	1	10	40	8	9

6 Conclusion

Principal-agent theory is an important innovative theory in informatization of internal control in colleges and universities. Based on it, this paper establishes a principal-agent model of informatization development model of internal control in colleges and universities. At the same time, it analyzes the relationship between the principal and the agent from the perspective of informationalized system development mode in colleges and universities. From theory and practice, the results show that compared to the "user development model", it is a better choice to completely allocate the task to the system development company. From the perspective of the informatization of internal control of institutions, the primary challenge faced by them is an effective connection between the existing information systems of various departments and the connection and sharing between the existing information systems and new information systems. Next, we will discuss the methods of effectively assigning internal control tasks for colleges and universities, as administrative institutions, when faced with realistic restraints. We also try to find out appropriate system development modes of informatization of internal control for these institutions. Hopefully, this paper can provide some theoretical support for informatization of internal control in administrative institutions.

References

1. Ministry of Finance, The Internal Control Standard for Administrative Institutions (Trial). 2012-11-29
2. Ministry of Education, The Guideline for Internal Control of Economic Activity of Colleges and Universities Directly of the Ministry of Education (Trial). 2016-4-20
3. Ministry of Finance, Outline of the 13th Five-Year Plan for Accounting Reform and Development. 2016-10-8
4. R. Marshall, P. John, Accounting information systems. Upper Saddle River **9**, 69–72 (2011)
5. M.-H. Yang, T.-L. Koo, The impact of information technology adoption on internal controls adaptation. Int J Bus Syst Res **08**(1), 14–33 (2014)
6. M. Popescu, S.-M. Popescu, G. Mangu, Internal management control approach by implementation of stages of the internal control management system in public entities. Intern Audit. Risk Manage. **42**(1), 53–58 (2016)
7. M. Huayang, T. Tong, Analysis of the internal control construction of the accounting information system under the internet of things environment. Commun. Financ. Acc. **28**, 102–103 (2013)
8. L. Guocheng, Y. Lili, Internal control management of meso-information systems coping with insider threat. Audit Econ. Study **6**, 49–55 (2014)
9. L. Juan, Research on enterprise accounting information system software development method selection based on AHP. China Manage. Inf. **17**(08), 38–40 (2014)
10. Z. Wei, Discussion on the choice of logistics information system development methods. Bus. Cult. (Acad. Ed.) **10**, 304 (2010)
11. B. Holmstrom, P. Milgrom, Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J. Law Econ. Organ. **7**(2), 24 (1991). https://doi.org/10.1093/jleo/7.special_issue.24

12. J. Zhen, *A Win-Win Cooperative Principal-Agent Model for Public-private Partnerships* (Chongqing University, 2016)
13. T. Hu, *Research on a Principal-Agent Model with the Organizational Structure* (Beijing Jiaotong University, 2016)
14. K. Hausken, Principal-agent theory, game theory, and the precautionary principle. Decis. Anal. **16**(2), 105–127 (2019). https://doi.org/10.1287/deca.2018.0380
15. H. Bi, F. Lu, S. Duan, M. Huang, J. Zhu, M. Liu, Two-level principal–agent model for schedule risk control of IT outsourcing project based on genetic algorithm. Eng. Appl. Artif. Intell. **91** (2020). https://doi.org/10.1016/j.engappai.2020.103584

Research on Balancing Problem of Stochastic Two-Sided Mixed-Model Assembly Lines

Beibei Zhang and Dan Chang

Abstract In the design of assembly lines in industrial systems, the adoption of two or more assembly lines is a method to realize high efficiency, which is called parallel assembly lines. In addition, the occurrence of stochastic events in industrial systems is inevitable, thus it is more realistic to regard industrial system as the stochastic environment. For this reason, this article proposes a mathematical model of stochastic mixed-model two-sided assembly line balancing problem (STMALBP), and adopts a genetic algorithm based on feasible sequence real-number encoding to solve STMALBP. The feasibility and effectiveness of the model and algorithm are verified by classical calculating examples, and the proposed algorithm is applied to balance the actual mixed-model assembly line with normal distribution operation time, which provides a valuable reference for the research of assembly line balancing.

Keywords Assembly line balancing problem · Genetic algorithm · Stochastic two-sided assembly line

1 Introduction

In industrial systems, efficient design of production units is crucial to productivity. When the demand is large enough, two-sided assembly lines can be used to improve assembly line productivity. Two-sided assembly line is a single line, of which operations can be implemented parallelly on both sides. It is widely adopted in the assembly of large and complex mechanical products such as automobiles, trucks and loaders [1]. Compared with one-sided assembly line, two-sided assembly line can shorten assembly line length, cut down equipment investment and maintenance costs, reduce material handling and improve the efficiency of workers. Therefore, it is of great significance to study the two-sided assembly line balancing problem

B. Zhang (✉) · D. Chang
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 19125510@bjtu.edu.cn

D. Chang
e-mail: dchang@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_66

(TALBP) to reduce the production cost and improve the competitiveness of enterprises. In addition, the characteristic of two-sided assembly line is the restriction of operation direction. Due to the combination of direction constraint and precedence constraint, the complexity of assembly line balancing problem is increased. There are two methods to solve TALBP: accurate algorithm and heuristic algorithm. Two-sided assembly line balancing problem is also a NP-hard problem. Therefore, researchers usually adopt heuristic algorithm to make TALBP close to the optimal solution, which usually meets the requirements of practical application. Kim et al. [2] proposed a meta-heuristic genetic algorithm to deal with TALBP with position constraint.

In addition, the occurrence of stochastic events in industrial systems is inevitable. Among these stochastic events that interfere with the production process, there are machine failures, operator unavailability, production line technical failures, complex tasks, emergencies in the internal logistics system, etc. All these reasons determine that the process time cannot be a constant, but a distribution fluctuating around a boundary value. From this perspective, the process time can be regarded as a stochastic variable. The assembly line balancing problem with uncertain operation time is generally solved by probability distribution of operation time or fuzzy time method. Aiming at the uncertain factors in the practical application of two-sided assembly lines, Özcan [3] established a stochastic constrained mixed integer programming model, which approximated TALBP by using stochastic operation time. Chiang et al. [4] improved this model by considering the probability of completing both sides of the workstation within the cycle time. Ozbakir and Tapkan [5] proposed a bees algorithm to deal with inaccurate targets by considering fuzzy multi-objective two-sided assembly lines. Tapkan et al. [6] dealt with the multi-objective two-sided assembly line balancing problem and solved the stochastic multi-objective two-sided assembly line balancing problem with three additional constraints. Chen et al. [7] proposed an improved ant colony algorithm to solve the balancing problem of two-sided mixed-model assembly lines, and obtained a task allocation scheme. Considering the concurrent operation, Yang [8] established a mathematical model of the two-sided mixed-model assembly line with concurrent operation constraints, and gave the solution process of genetic algorithm. On the premise of considering the balance characteristics of two-sided assembly lines, Song [9] proposed a heuristic algorithm for task priority allocation with operation direction constraints to obtain different balance schemes and satisfying results. Yu et al. [10] analyzed the uncertainty factors in the uncertainty problem existing in the hybrid assembly line based on the mathematical model, and adopted eM-Plant modeling to study the influence of uncertainty factors on the balance of the hybrid assembly line through system simulation. Zhang et al. [11] considered the characteristics of stochastic mixed-model assembly lines and proposed a hybrid particle swarm optimization algorithm to solve the balancing problem of stochastic mixed-model assembly lines. Huang et al. [12] put forward an algorithm based on beam search to solve the stochastic mixed-model assembly line balancing problem.

The balancing problem of two-sided mixed-model assembly line exists widely in actual production of enterprises, especially the influence of the uncertainty of operation time on the balance of two-sided mixed-model assembly line cannot be ignored [13–15]. Therefore, the research on stochastic two-sided mixed-model assembly line balancing problem (STMALBP) has more practical significance and research value [16–18]. In this article, the stochastic manufacturing of assembly operation is taken into account, STMALBP model is constructed, and genetic algorithm is adopted to solve the balancing problem. Finally, a specific balancing scheme is given to solve the balancing problem existing in the assembly line D of Company Y, which verifies the effectiveness of the model and algorithm.

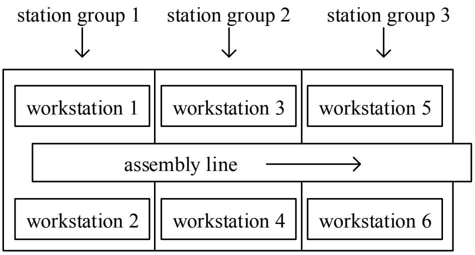
2 Analysis on Balancing Problem of Stochastic Mixed-Model Assembly Line

2.1 Complexity of Two-Sided Assembly Line Problem

In the two-sided assembly line, both sides can complete different tasks in parallel. The opposite stations are called paired stations, and the opposite stations can also be called accompanying stations, as workstation 1 and workstation 2 shown in Fig. 1.

Two-sided assembly lines are more complicated than one-sided assembly lines because each operation element in the two-sided assembly lines has a specified operation direction attribute. In addition to the limitation of operation direction, the operation elements on the left and right workstations of the two-sided assembly lines will generate a "waiting time" due to the restriction of the operation sequence, while the size of this "waiting time" is closely related to the operation sequence of the operation elements. As shown in Fig. 2, for a certain workstation group in the two-sided assembly line, the operation elements assigned to the two workstations in the same workstation group may cause inevitable "waiting time" due to the constraint of the operation sequence relationship. Thus, when calculating the workstation operation time in the two-sided assembly line balancing problem, "waiting time" should be taken as part of the workstation operation time.

Fig. 1 Layout of two-sided assembly line



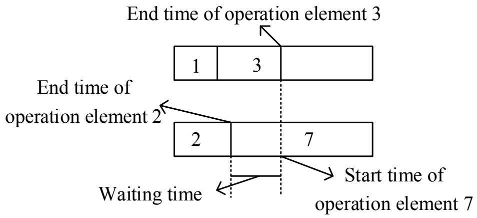

End time of operation element 3
1
3
End time of operation element 2
2
7
Waiting time
Start time of operation element 7

Fig. 2 Description of waiting time

2.2 The Assembly Process Being Affected by Stochastic Factors

In previous studies on the assembly line balancing problem, the assembly operation time was generally considered as constant, which is called the deterministic assembly line balancing problem. The deterministic assembly line balancing problem is described as the average production cycle $C = T/D$ in a planning period T, where T is the effective production time in the planning period and D is the demand in the planning period. Given that each operation element in the product assembly is I (process), the set of operation elements (process set) is P, and the precedence relation between each operation element is given, the exact time of the i-th operation element time in the production of the product is t_i (i = 1, 2, ..., I) through operation measurement. However, under the influence of various factors in the real industrial system, the working time of the operators changes stochastically, which can be described by stochastic distribution function. The corresponding assembly line balancing problem with stochastic working time is stochastic assembly line balancing problem (SALBP). Since the influence of stochastic factors on workstation time can be converted into the influence on process time, and the influence of stochastic factors on workstation time obeys normal distribution, then the influence assigned to each process will obey normal distribution, so $t_i \sim N(\mu_i, \sigma_i^2)$.

2.3 Quick Response to Market Requirements

With the improvement of modern living standards, people's demand for products is becoming more and more diversified, and the single production mode is gradually replaced by multi-variety and small-batch mixed-model production mode. Under the mixed-model production mode, the types and quantities of products can change according to the changes of the market and customer needs, which can quickly respond to the market and meet the personalized needs of different consumers. Based on the change of market demand in the production process, this article assumes that the demand D obeys Poisson distribution.

Mixed-model production is carried out on two-sided assembly lines and various stochastic factors are considered, then we can call this type of research object stochastic mixed-model assembly lines. Today's market places more emphasis on "flexibility" and is customer-oriented. Traditional simple assembly lines are difficult to respond quickly. Therefore, this article aims to study the stochastic mixed-model assembly line balancing problem (SMALBP). A mathematical model of type II stochastic mixed-model two-sided assembly line balancing problem (STMALBP-II) is established, which aims at minimizing the production cycle. A genetic algorithm based on feasible sequence real number encoding is adopted to solve STMALBP-II. Finally, the balance model is used to optimize the balancing of the mixed-model assembly line D of Company Y, which solves the balancing problem existing in the assembly line.

3 Construction of Assembly Line Balancing Model Based on Genetic Algorithm

3.1 Determination of STMALBP Variables and Constraints

When the operation time on the two-sided assembly line is a stochastic variable, the time of the whole workstation becomes a stochastic variable due to the stochasticity of the operation time. Assuming that the actual operation time of each operation element obeys normal distribution, namely $t_i \sim N\left(u_i, \sigma_i^2\right)$. For the convenience of expression, the relevant variables are listed:

P—task set, $P = \{1, 2, \ldots, \mathrm{i}\}$;

$P_i = \{t_i, d\}$—the i-th process, each process contains the two attributes of time and direction;

P_E—task set that can be assigned to both the left assembly line and the right assembly line, $P_E \in P$;

P_L, P_R—task sets assigned to each station of the assembly line on the left and right sides, $P_L, P_R \in P$;

u_{ni}, σ_{ni}^2—expected values and variances of average operation time of the i-th process of type N product respectively;

P_J—the set of tasks assigned to station j;

T_j, T_j'—the total operating time of station j and station j';

E_j—station j' utilization rate;

R_j, R_j'—the remaining time of station j and station j';

$P(i)$—all preorder task sets of task i;

B_j—Boolean variable, indicating whether station j is enabled or not, if P_J is not null, then $B_j = 1$, otherwise it is 0;

d—operation direction constraints, $d \in \{L, R, E\}$;

S—stochastic variables with values belonging to interval (0, 1);

x_{nik}—decision variable, k is the serial number of the station assigned to task I of the type n product;

x_{wjd}—represents the number of tasks assigned to position *w* and station *j* with operation direction *d*.

Stochastic two-sided mixed-model assembly lines have not only the constraints of process indivisibility and task sequence relation existing in general assembly lines, but also the constraints of operation direction and sequence-related completion time of tasks. Therefore, the balance model must meet the following constraints:

(a) *Processes cannot be divided, but can and can only be assigned to one station*

$$\begin{cases} P_j \cap S_l = \emptyset \\ \bigcup_{k=1}^{m} P_k = I, k = 1, 2, \ldots, M \end{cases} \tag{1}$$

(b) *Precedence relation constraint*

$$\sum_{k=1}^{m} (kx_{njk} - kx_{nik}) \geq 0 \tag{2}$$

where i is the immediate predecessor activity, or the preorder process of j.

(c) *The relationship between the demands of various varieties*

$$\sum_{n=1}^{N} q_n = 1 \tag{3}$$

(d) *Any station time plus waiting time of any product is less than the cycle*

$$\begin{aligned} \max(T_{nk}) + \sum_{i=1}^{I} x_{nik} D_{nik} &= \max\left(\sum_{i=1}^{I} t_{ni} x_{nik}\right) + \sum_{i=1}^{I} x_{nik} D_{nik} \\ &= \sum_{i=1}^{I} (x_{nik} D_{nik} + u_{ni} x_{nik}) \leq C \end{aligned} \tag{4}$$

Assuming $t_{ni} \sim N\left(\mu_{ni}, \sigma_{ni}^2\right)$, D_{nik} represents the waiting time before the operation of process i when process i of n-th type product is assigned to station K, while there is the preorder process of process i and its operation completion time is later than the operation start time of process *i*, $n = 1, 2, \ldots, N$; $i = 1, 2, \ldots, I$; $k = 1, 2, \ldots, K$.

(e) *Average cycle*

$$\overline{C} = T/D = \sum_{i=1}^{I} \sum_{n=1}^{N} q_n t_{ni} x_{nik} = \sum_{i=1}^{I} \sum_{n=1}^{N} \lambda_n \mu_{ni} x_{nik} \tag{5}$$

$\overline{C}$ is the average production time for n-th type product at station k

(f) *Direction constraint*

$$x_{ni} mod 2 = \begin{cases} 0, k \in I_R \\ 1, k \in I_L \end{cases} \tag{6}$$

This article studies the assignment of all operation elements to the corresponding workstations when the number of workstations is fixed, so as to minimize the cycle time, thus the objective function is Min C.

3.2 STMALBP Balancing Model Construction

Stochastic two-sided mixed-model assembly line balancing problem class II (STMALBP-II) is a combinatorial optimization problem according to some constraints and related objectives. When the scale of the problem is small, accurate methods (such as branch and bound method, Lingo software, etc.) can be used to obtain the optimal solution in a short time. However, with the increase of the scale of the problem, the time and difficulty of solving the problem will greatly increase. This kind of algorithms often consume a lot of computational resources, and sometimes it is even impossible to obtain the optimal solution, which is a typical NP-hard problem. In this case, a more practical idea is to adopt heuristic method to obtain the approximate optimal solution. Similar to other combinatorial optimization problems, the common heuristic optimization algorithms for solving this problem include tabu search, simulated annealing and genetic algorithm [9–11]. In view of the satisfying effect of genetic algorithm in solving NP-hard problem, this algorithm is selected to solve STMALBP-II.

For the convenient expression of the solution process of the whole genetic algorithm, this chapter will describe each step of the genetic algorithm with examples. The precedence relation diagram of the examples is shown in Fig. 3.

1. *Encoding*

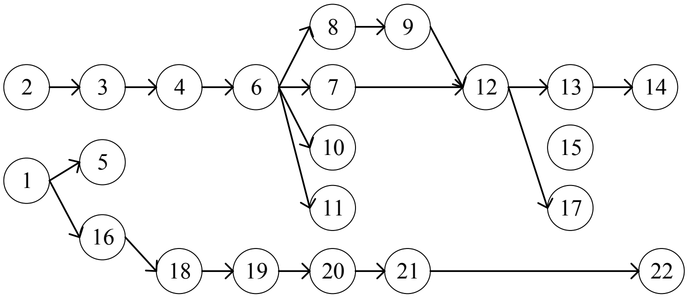


Fig. 3 Precedence relation example diagram

1	5	2	3	4	16	6	8	11	9	10	18	19	20	21	7	12	13	17	14	22	15
1	1	0	0	0	1	1	1	0	0	1	1	1	1	1	0	0	0	1	0	0	1

Fig. 4 lustration of a chromosome

Unlike one-sided assembly lines, the balancing of two-sided assembly lines is not only restricted by the precedence relation, but also by direction constraint, when calculating the station time, waiting time must be considered. In this article, a real number encoding method of genetic algorithm based on feasible task sequence is proposed basing on the precedence relation diagram of corresponding processes to solve assembly line balancing problem, which can ensure that the optimal solution search is only carried out in the subspace of feasible task sequence with high efficiency.

In this article, the encoding is different from that of the traditional gene expression. The traditional gene expression is usually a 1 * n matrix, namely, a chromosome has only one row for concrete expression. In this article, a 2 * n matrix is used to represent the encoding. The first row of the matrix represents the operation element, namely, the process; the second row represents the direction constraint of the corresponding process, with values of 0 (the process is actually assigned to the right side) and 1 (the process is actually assigned to the left side). This encoding method is easier to read than the traditional representation method, and is also convenient for computer calculation. And it is carried out on the premise of meeting the requirements of precedence relation, so it is also an encoding method based on feasible solution. An encoding for Fig. 3 is shown in Fig. 4.

There are various encoding methods that satisfy the precedence relation, which makes it possible to generate a certain number of initial populations and seek the optimal solution through crossover and mutation.

2. *Generating initial population*

The generation of the initial population adopts the method of stochastic generation. That is, on the basis of encoding, each encoding obtains different chromosomes, so the generated chromosomes are stored in a matrix to form the initial population, the size of the initial population can be set according to the needs of the research.

3. *Decoding*

Decoding is the process of converting from genotype to phenotype, in other words, it is a process of solving problems by converting abstract codes into actual objects. In this article, genetic algorithm is used to solve the balancing problem of stochastic mixed-model two-sided assembly lines. Decoding is to decode the initial population generated by the stochastic mixed-model two-sided assembly line on the premise of the previous encoding step.

The first step is to calculate the lower limit of the cycle according to the existing conditions. Because in actual production, there are various uncertain factors and other related factors, so that the actual production cycle cannot be simply equal to the sum of the time of each process divided by the number of workstations; Even if the influence of direction, the sum of the total process time on the left divided by the number of workstations on the left and the sum of the total process time on the right divided by the number of workstations on the right are taken into account, and the larger of the two is taken as the cycle, it cannot meet the production requirements. Another basic constraint is that the cycle must be greater than the maximum process time. Therefore, the lower limit of cycle is obtained as follows:

$$C_{\min} = \max\left\{\left|\frac{T_{sum}}{k}\right|, \left|\frac{T_{suml}}{\frac{k}{2}}\right|, \left|\frac{T_{sumr}}{\frac{k}{2}}\right| \max(t_{ni})\right\} \tag{7}$$

$$T_{sum} = \sum_{i=1}^{I}\sum_{n=1}^{N} t_{ni} D_n \tag{8}$$

$$T_{sumr} = T_{suml}$$

$$\begin{cases} T_{sumr} = \sum\limits_{i=1}^{I}\sum\limits_{n=1}^{N} t_{ni} D_n \\ P(2, i) = 0 \end{cases} \tag{9}$$

$$\begin{cases} T_{suml} = \sum\limits_{i=1}^{I}\sum\limits_{n=1}^{N} t_{ni} D_n \\ P(2, i) = 1 \end{cases} \tag{10}$$

where k is the number of workstations, T_{sum} is the sum of all process times, T_{suml} is the sum of all process times on the left, $\mathrm{T_{sumr}}$ is the sum of all process times on the right, t_{ni} is the time of process i of the n-th product, and D_n is the demand for the n-th product. $P(2, i)$ is direction constraint, $P(2, i) = 0$ represents the right and $P(2, i) = 1$ means the left.

The second step is to allocate the processes. Before allocation, the direction must be judged. The allocation starts from the first position one by one, and the process is also allocated in the encoding order. For each process allocated, the workstation time of the workstation must be calculated. If the current workstation time plus the process time to be allocated is less than the lower limit of the cycle, the process to be allocated can continue to be allocated to the workstation, otherwise, the next workstation needs to be opened. After opening the workstation, it is necessary to first judge whether it is the last one, and if so, all the remaining processes in the same direction will be assigned to the workstation.

The third step is to calculate the cycle time. After all processes are assigned, the current cycle is calculated. Comparing the current cycle with the lower limit of the cycle, if the current cycle is greater than the lower limit of the cycle, the lower limit of the cycle needs to increase by a certain step size, then return to Step 2.

The fourth step is recycling. The second and third steps are recycled until the cycle equals the lower limit of the cycle, then decoding is complete.

4. *Fitness function and selection operation*

After the assembly line is balanced, it is necessary to evaluate the advantages and disadvantages before and after the balancing. Therefore, the balance rate of the assembly line is taken as the evaluation function of chromosome (11). The greater the P value, the better the chromosome and the better the corresponding balance scheme. The roulette wheel selection is used for selection and elimination.

$$P = \frac{\sum_{k=1}^{m} T(S_k)}{m \times \max[T(S_k)]} \times 100\% \tag{11}$$

5. *Crossover and mutation*

(a) In this article, stochastic and simple crossover is used to carry out crossover operation. Examples are combined to generate stochastic male parents P1 and P2, stochastic number k = 7, and progenies C1 and C2 are generated by crossover. The situation of C1 and C2 before and after crossover is shown in Figs. 5 and 6.

(b) Mutation is to exchange some genes in the chromosome to form a new chromosome, thus increasing the number of feasible solutions. Figure 7 is a schematic diagram to show the mutation operation.

6. *Termination Conditions*

P1	2	3	1	16	5	4	6	11	8	9	10	18	7	12	17	19	20	21	15	22	13	14
	0	0	1	1	1	0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0

↕

P2	1	5	2	16	3	18	4	19	20	6	10	21	7	8	9	11	12	15	13	17	22	14
	0	1	0	1	0	1	0	1	1	0	1	1	0	1	0	0	0	1	0	1	0	0

Fig. 5 The crossover of male parents P1 and P2

C1	2	3	1	16	5	4	18	19	20	6	10	21	7	8	9	11	12	15	13	17	22	14
	0	0	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	1	0	1	0	0

C2	1	5	2	16	3	18	6	11	8	9	10	7	12	17	19	20	21	15	22	13	14	4
	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	1	0	1	0	1	0	0

Fig. 6 Progenies C1 and C2 obtained after crossover

P1	2	3	1	16	5	4	6	11	8	9	10	18	7	12	17	19	20	21	15	22	13	14
	0	0	1	1	1	0	0	0	1	0	1	1	0	0	1	1	1	1	1	0	0	0

↕

C1	2	3	1	16	5	4	18	19	20	6	10	21	7	8	9	11	12	15	13	17	22	14
	0	0	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	1	0	1	0	0

Fig. 7 Schematic diagram of chromosome before and after mutation

Generally, the algorithm terminates when the fitness of the optimal individual and the population no longer rise, or when the current evolutionary algebra reaches the preset algebra. In this article, when the current evolutionary algebra reaches the preset maximum genetic algebra, the genetic algorithm stops.

3.3 Balancing Model Realization and Effect Verification Based on MATLAB

1. *Realization process of balancing optimization model*

According to the flow of genetic algorithm, this article uses MATLAB to establish a balancing optimization model. The functions used to establish the model are shown in Table 1. Firstly, the parameters are defined in the main function Main. The initial population N = 100, the crossover probability Pc1 = 0. 6, the mutation probability Pm = 0. 15, and the evolutionary algebra N (G) = 150. After defining the parameters, the number of workstations, workstation time and precedence relation matrix are taken as input quantities, and different functions are called to complete the establishment of the balancing optimization model. The modeling process is shown in Fig. 8.

2. *Verification of the effect of balancing optimization model*

After the balancing optimization model is established, it is necessary to verify the effectiveness of the model. In this article, the effectiveness of the model is verified by the classical Jackson assembly line balancing problem. Jackson process priority diagram is shown in Fig. 9. Jackson balancing problem adopts enumeration method to balance and optimize the production line, and the time of each process is Time = [6 2 5 7 1 2 3 6 5 5 4]. The balancing scheme for Jackson balancing problem is obtained through the optimization by the established balancing optimization model, and then the balancing optimization results of genetic algorithm are compared with the traditional method to solve Jackson problem—enumeration method, and the comparison results are shown in Table 2. Through the two evaluation indexes in

Table 1 MATLAB functions correspondence table of genetic algorithm

No.	Function name	MATLAB function name
1	Main function	Main.m
2	Initialization function	Ini_Pop.m
3	Decoding function	Decoding.m
4	Fitness function	Fitness.m
5	Select function	Selection.m
6	Crossover function	Crossover.m
7	Mutation function	Mutation.m
8	Drawing function	Draw.m

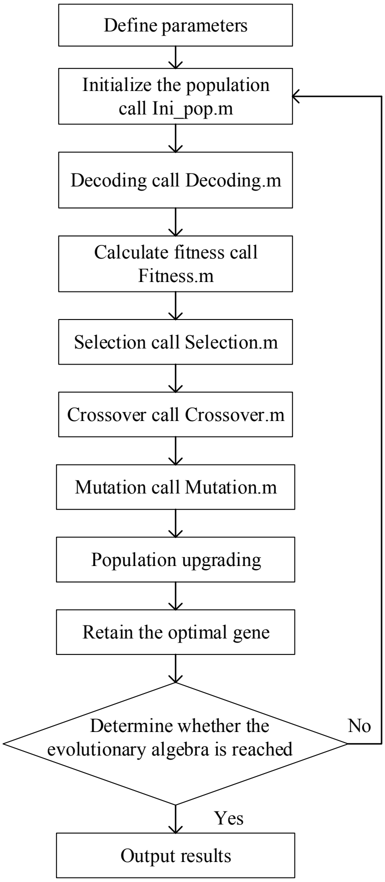

Define parameters
Initialize the population call Ini_pop.m
Decoding call Decoding.m
Calculate fitness call Fitness.m
Selection call Selection.m
Crossover call Crossover.m
Mutation call Mutation.m
Population upgrading
Retain the optimal gene
Determine whether the evolutionary algebra is reached
No
Yes
Output results

Fig. 8 Flow chart of the realization of MATLAB balancing optimization model

Fig. 9 Jackson process priority diagram

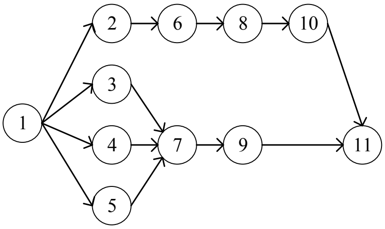


Table 2 Comparison of Jackson balancing optimization effect

Method	Workstation						
	1	2	3	4	5	P (%)	SI
Jackson Enumeration method	10	7	10	10	9	92	3.16
Genetic algorithm	9	8	10	10	9	92	2.50

the table—production line balance rate P and balance index SI, it can be seen that the production line balance rate P optimized by genetic algorithm and enumeration method is the same, while the smoothness index SI of genetic algorithm is less than enumeration method, so the optimization scheme of genetic algorithm is better, and the established balancing optimization model is effective.

4 Research on Application of Stochastic Mixed-Model Two-Sided Assembly Line

4.1 Analysis of Assembly Status of Company Y

In this article, the assembly line D of Company Y is taken as the research object. It is known that the assembly line currently has 5 positions and 9 workstations, of which 1 position and 2 workstations often appear imbalance, that is, bottleneck workstations are these 2. Because the whole assembly line D is not balanced, it often leads to line shutdown and low productivity. Assembly line D belongs to mixed-model assembly line and produces two varieties A and B. The demand for each variety is uncertain. Assuming that the production cycle is one month, it obeys Poisson distribution with lamda_A = 300 and lamda_B = 200 respectively. The workstation time of these 9 workstations follows a normal distribution with a mean value of 186 and a variance of 15.

According to the process precedence relation diagram of products A and B, a combined precedence relation diagram is drawn, as shown in Fig. 10. The process time of various products of assembly line D is shown in Table 3).

MATLAB is used to stochastically generate the demand for varieties A and B, as shown in Table 4.

Because the research object is mixed-model assembly line, and the uncertainty of demand and the process time of each variety are different, the comprehensive process time should be calculated to convert stochasticity into certainty so as to facilitate the follow-up work.

According to the actual situation, the calculation formula of comprehensive process time is as follows:

$$t_i = \sum_{m=1}^{M} \sum_{n=1}^{N} t_{ni} D_{mn} / \sum_{n=1}^{N} D_{mn} \tag{12}$$

where i is the process number, and its value range is i ∈ [1, 27]; n is the product variety, and its value range is n ∈ [1, 2]; m is the m-th calculation, and its value range is m ∈ [1, 30]. After calculation, the comprehensive process time can be obtained as shown in Table 5.

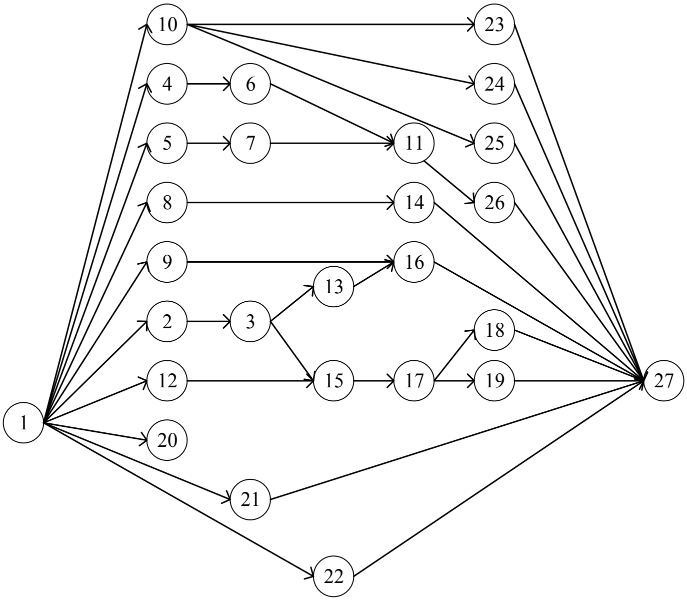


Fig. 10 Combined precedence relation diagram

Table 3 Process time of various products in assembly line D

Process	Operation time (s)		Process	Operation time (s)	
	t_{Ai}	t_{Bi}		t_{Ai}	t_{Bi}
1	50.5	53.4	15	120	122.7
2	121	124.6	16	93	93
3	91.5	93.3	17	118.8	120.3
4	127	125	18	11	10.9
5	10.9	12.1	19	45.3	39.7
6	43	46.3	20	16.5	15.5
7	18.7	20.4	21	64	63
8	62	64	22	36.7	38.6
9	116.2	128	23	87	80
10	29	34.8	24	36.8	40.2
11	35	43.5	25	45	44.7
12	44.9	50.3	26	80.1	79.4
13	76	78	27	32	29.5
14	53.1	53			

Table 4 Demand for varieties A and B in 30 production cycles

Variety No.	A	B	Variety No.	A	B	Variety No.	A	B
1	298	195	11	267	212	21	295	202
2	293	221	12	261	175	22	268	199
3	288	170	13	327	202	23	336	213
4	289	196	14	297	202	24	320	184
5	328	178	15	307	187	25	317	201
6	281	185	16	305	178	26	283	214
7	313	201	17	315	206	27	329	227
8	329	217	18	309	185	28	295	208
9	296	224	19	292	196	29	307	198
10	302	189	20	285	216	30	294	194

Table 5 Comprehensive process time

Process	Process time (s)	Process	Process time (s)	Process	Process time (s)
1	52.1	10	31.5	19	42.3
2	123.2	11	38.2	20	15.9
3	93.7	12	47.8	21	63.2
4	125.9	13	77.1	22	37.4
5	10.8	14	53.4	23	84.1
6	45.3	15	120.6	24	37.8
7	19.4	16	93.2	25	44.6
8	62.6	17	118.7	26	79.3
9	120.3	18	11.4	27	30.2

4.2 Optimization of Assembly Line Balancing

MATLAB (2018a) simulation software and genetic algorithm are applied to solve the problem to carry out balance optimization on the mixed-model assembly line. Among them, the population iterations are 150 times, the crossover probability is 0.6, and the mutation probability is 0.15. After 150 iterations, the optimization results are shown in Fig. 11. The process distribution of the 9 workstations tends to be balanced and the production cycle is optimized. The processes on the 9 stations have been redistributed, and the distribution of processes is shown in Table 6.

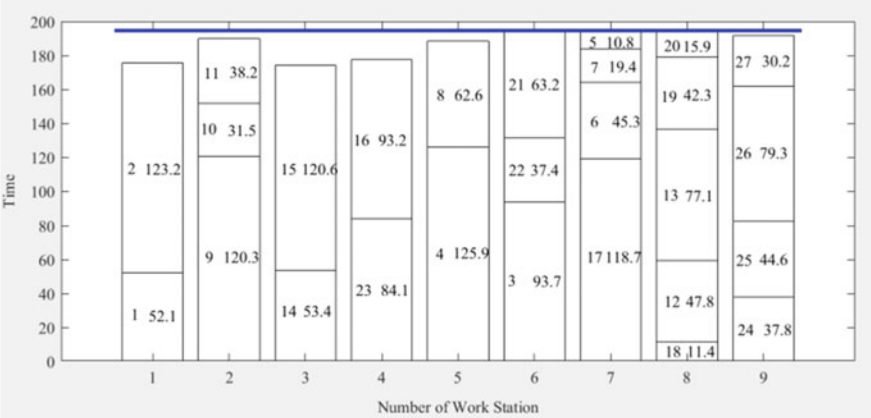


Fig. 11 Task allocation diagram

Table 6 Allocation scheme of optimized assembly line workstations

Position	Workstation	Station time	Process assigned to this station
1	1	175.3	1, 2
	7	194.2	17, 6, 7, 5
2	2	188.5	9, 10, 11
	5	188.5	4, 8
3	3	173	14, 15
	6	194.3	3, 22, 21
4	4	177.3	23, 16
	8	194.5	18, 12, 13, 19, 20
5	9	191.9	24, 25, 26, 27

4.3 Analysis of Results

Comparing the station time before and after optimization of the assembly line D of Company Y, the production cycle of the assembly line D after optimization is reduced, and the time distribution of each station is more balanced, as shown in Fig. 12.

The balance optimization effect of assembly line D of company Y can be expressed by the change of balance rate and production cycle, where in the balance rate P of assembly line is obtained by Formula (13):

$$P = \frac{\sum_{k=1}^{m} T(S_k)}{m \times \max[T(S_k)]} \times 100\% \tag{13}$$

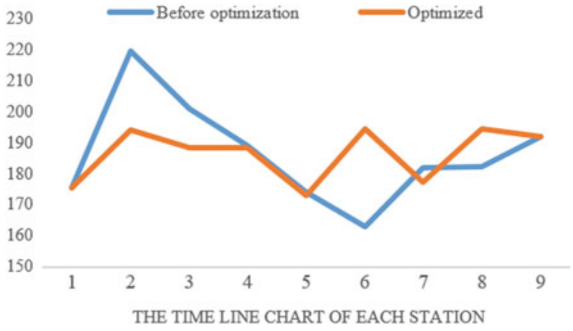


Fig. 12 The time line chart of the station before and after optimization

Table 7 Comparison of optimization effects

Evaluation index	Cycle time (s)	Balance rate P (%)
Before optimization	219.6	84.9
After optimization	194.5	95.8

According to Fig. 11 and Table 7, the production time of each station after optimization is relatively balanced, and the time of each station is always below 200 s, with the original maximum workstation time, namely, the cycle time being 219.6 s; The optimized maximum workstation time, namely, the cycle time, decreased to 194.5 s, and the optimized assembly line balance rate increased from 84.9 to 95.8%, which is effectively optimized.

5 Conclusions and Prospects

In actual production, two-sided assembly line are well adopted in the production of most large equipment and machinery. In order to provide some theoretical basis for the balancing problem of two-sided assembly lines in actual production, this article takes assembly line D of automobile company Y as the research object to find the optimal solution of STMALBP-II. The influence of various stochastic factors on assembly line balancing can be converted into their influence on process time, which obeys normal distribution. Meanwhile, customer demand is taken as the direction, the demand of each variety in the mixed-model assembly line studied in this article is stochastic. A genetic algorithm based on feasible sequence is designed to solve the STMALBP-II, and the feasibility of the algorithm is proved by calculating examples.

This article studies STMALBP on a certain premise. For instance, the design of workstations on both sides of the two-sided assembly line is symmetrical, and the specific number of workstations to be opened depends on the actual situation. In actual production, there may be some processes that do not require two-sided operation, so there may be only one workstation in a certain position of the line.

The above-mentioned situation needs to be further studied, and the corresponding mathematical models also require improvement.

Acknowledgements This work was partially supported by Beijing Logistics Informatics Research Base and Intelligent Emergency Logistics Linkage System of Public Emergencies in Beijing under Grant 18JDGLB019.

References

1. Z. Li, Q. Tang, L. Zhang, Two-sided assembly line balancing problem of type I: improved, a simple algorithm and a comprehensive study. Comput. Oper. Res. 79 (2017)
2. Y.K. Kim, Y. Kim, Y.J. Kim, Two-sided assembly line balancing: a genetic algorithm approach. Prod. Plan. Control **11**(1) (2000)
3. U. Özcan, Balancing stochastic two-sided assembly lines: a chance-constrained, piecewise-line, mixed integer program and a simulated annealing algorithm. Eur. J. Oper. Res. **205**(1) (2009)
4. W.-C. Chiang, T.L. Urban, C. Luo, Balancing stochastic two-sided assembly lines. Int. J. Prod. Res. **54**(20), 6232–6250 (2016)
5. L. Ozbakir, P. Tapkan, Balancing fuzzy multi-objective two-sided assembly lines via Bees algorithm. J. Intell. Fuzzy Syst.: Appl. Eng. Technol. **21**(5), 317–329 (2010)
6. A. Baykasolu, P. Tapkan, L. Özbakir, Bees algorithm for constrained fuzzy multi-objective two-sided assembly line balancing problem. Optim. Lett. **6**(6), 1039–1049 (2012)
7. J. Chen, Q. Zhang, Ant colony algorithm in the application of the assembly line balancing problem. Comput. Era **2008**(12), 20–22 (2008)
8. T. Yang, L. Jiansha, L. Kong, Study on balancing problem of two-sided assembly line with concurrent operation. J. Zhe Jiang Univ. Technol. **39**(4), 440–444 (2011)
9. L. Song, Z. Zhang, W. Chen, Heuristic for two-sided stochastic assembly line balancing. Ind. Eng. J. **14**(4), 129–134 (2011)
10. Y. Zhaoqin, Simulation of uncertainty for mixed-model assembly line balancing problem. China Mech. Eng. **19**(11), 297–1302 (2008)
11. Z. Zhang, Y. Qingliang, Hu Junyi, Hybrid particle swarm optimization algorithm for balancing problem of stochastic mixed-model assembly line. Mach. Des. Res. **29**(2), 60–63 (2013)
12. W. Huang, J. Zhang, Z. Zhou, A beam search approach to stochastic mixed-model assembly line balancing problem. Oper. Res. Manage. Sci. **19**(6), 20–26 (2010)
13. M.S. Yang, L. Ba, H.Y. Zheng, Y. Liu, X.F. Wang, J.Z. He, Y. Li, An integrated system for scheduling of processing and assembly operations with fuzzy operation time and fuzzy delivery time. Adv. Prod. Eng. Manage. **14**(3), 367–378 (2019)
14. C. Jiang, J.T. Xi, Dynamic scheduling in the engineer-to-order (ETO) assembly process by the combined immune algorithm and simulated annealing method. Adv. Prod. Eng. Manage. **14**(3), 271–283 (2019)
15. S.L. Yang, Z.G. Xu, J.Y. Wang, Modelling and production configuration optimization for an assembly shop. Int. J. Simul. Model. **18**(2), 366–377 (2019)
16. H. Wu, A. Tsai, H. Wu, A hybrid multi-criteria decision analysis approach for environmental performance evaluation: an example of the TFT-LCD manufacturers in Taiwan. Environ. Eng. Manage. J. **18**, 597–616 (2019)
17. G.I. Fragapane, C. Zhang, F. Sgarbossa, J.O. Strandhagen, An agent-based simulation approach to model hospital logistics. Int. J. Simul. Model. **18**(4), 654–665 (2019)
18. M.S. Yang, L. Ba, Y. Liu, H.Y. Zheng, J.T. Yan, X.Q. Gao, J.M. Xiao, An improved genetic simulated annealing algorithm for stochastic two-sided assembly line balancing problem. Int. J. Simul. Model. **18**(1), 175–186 (2019)

Intralogistics Conversion of the Physical Internet

Eszter Puskás and Gábor Bohács

Abstract In our previous research, we have identified the components of the Physical Internet (PI, π) that are required to build and operate a Physical Internet based network. The components we defined include specific physical units, such as the π-container or the π-hub. In addition, components of management and information systems were included such as Open monitoring system or Track and Trace system. In this study we are exploring the feasibility of converting the established system of Physical Internet. Each component will be individually examined to find out which component is suitable for the PI network. The study discusses the applicability of existing intralogistics to a Physical Internet based network. In conclusion, we are exploring the possibility of future concepts for the intralogistics conversion of the Physical Internet.

Keywords Physical Internet (PI) · Intralogistics · PI components · Simulation

1 Introduction

Nowadays, global logistics and supply chain management practices are unsustainable. Therefore, there is a clear need in the world to create an economically, environmentally and socially sustainable system [1]. In this respect, the importance of intralogistics systems, which can be defined as part of the supply chain, should not be neglected either, as growing individual needs also create an increasing challenge to the high quality of these systems. Given that intralogistics is a key cost factor for many companies, it can be seen as the focus of operational logistics activities [2].

Montreuil identified 13 specific symptoms of supply chain management practices that are a "global challenge to global logistics sustainability" [3]. For the identified

E. Puskás (✉) · G. Bohács
Department of Material Handling and Logistics Systems, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: eszter.puskas@logisztika.bme.hu

G. Bohács
e-mail: gabor.bohacs@logisztika.bme.hu

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_67

problem, he proposed a comprehensive review of the logistics sector and defined the foundations of the concept of the Physical Internet (PI, π), which is based on the metaphor of the digital Internet [3]. The concept fundamentally breaks with traditional storage, transport and movement solutions, creating a holistic concept based on the connectivity of the physical, digital and operational worlds. The implementation of PI requires a change of approach and the development of new methods and tools in many areas. In recent years, more and more attention has been paid by both academics and practitioners to exploring the PI concept and developing the solutions and models available to it. This is also shown by the growing trend in PI publications and investments in PI-related projects in the international literature [4].

In our previous research, we have identified the components of the PI concept that include elements that perform both physical and IT and management tasks [5]. In our article, we would like to define the components of a Physical Internet-based intralogistics system based on the previously developed systematics. To do this, we examine the intralogistics suitability and possibilities of the elements defined for supply chains one by one.

Examining the relationship between Industry 4.0 and the Physical Internet, we can conclude that among the tools of Industry 4.0, the simulation methodology should be applied to model PI-based systems [5]. Therefore, in this study, we use a simulation method to create the proposed physical Internet-based intralogistics system.

This paper was structured as follows. This section briefly introduces the topic we have focused on. In the second section, we present the concept of the Physical Internet. The third section discusses the components of the Physical Internet along with defining its intralogistics conversion. In the fourth section, we introduce the Physical Internet-based intralogistics system's simulation model by creating a discrete-event simulation. Finally, in the last chapter, we summarize the results of the article and identify further research directions.

2 The Physical Internet Concept

Decades ago, the emergence of the digital world was inspired by the physical world. The principle of the Physical Internet is based on the metaphor of the digital internet. In the digital world, data is transmitted according to the TCIP/IP protocol, using data packets, encrypting and making information properly available. The PI wants to apply this principle to the management of physical objects. In a globally collaborative logistics network, participants work together and cooperate with each other to perform their operational tasks using a unified equipment and communication system.

The foundations of the physical internet (PI, π) were formulated by Montreuil [3]. The PI concept is an open, global logistics system. Its main goal is to create a sustainable system that can operate effectively worldwide. The expected efficiency is achieved by optimal operation of the interconnection system between the network elements. The global vision requires operation within a standard framework, the principles and requirements of which should be followed for PI participants in the logistics network. This logistics system is based on a set of eight foundations as shown in the Fig. 1. The first is to create efficiency and sustainability from a social, economic and environmental point of view. The second pillar is the development of an open, interconnected Logistics Web. The combination of the previous two determines the third basis, which is to create an open global logistics system. The fourth fund, universal interconnectivity, is the key to the defined future system. The fifth basis is the encapsulation required to create physical packets, modeled on the digital Internet. The sixth base includes the interfaces, while the seventh fixes the importance of the basic protocols. Control by technological, business and infrastructural innovation corresponds to the eighth base.

The implementation of PI requires a change of approach and the development of new methods and tools in many areas. For example, the experts examined the issue of new cost models, inventory, and the possibilities of the Physical Internet in city-logistics [4]. A number of studies have been conducted to examine the elements and functioning of the supply chain. However, there are very few articles on Physical Internet-based intralogistics systems.

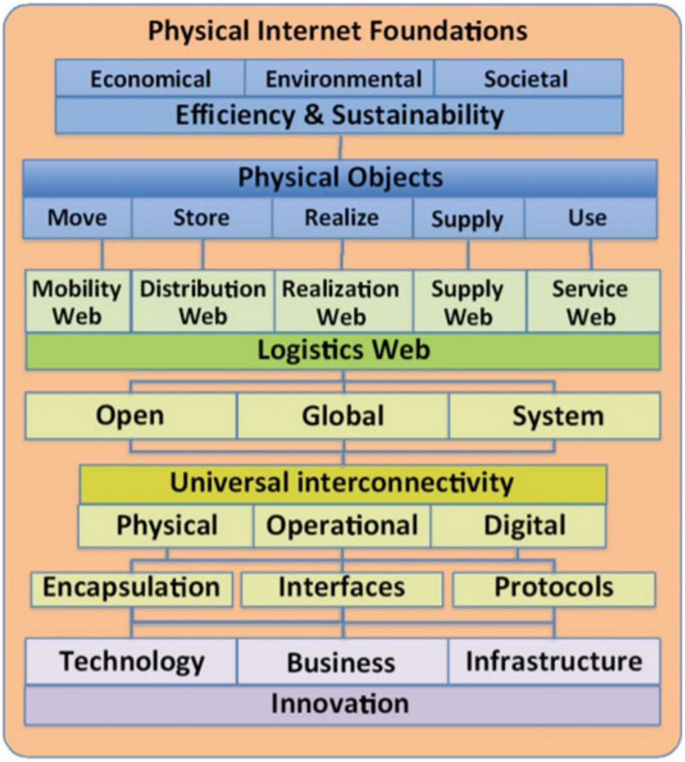


Fig. 1 Fondations of the Physical Internet [3]

3 Components of the Physical Internet-Based Intralogistics System

In this chapter, we present an intralogistics conversion of the Physical Internet by examining each component individually.

In our previous research, we created a new systematics to define the components of the Physical Internet [5]. A new value in the systematics is that in addition to the physical units that create the system, such as the π-container, it also contains components that perform information technology (IT) and management tasks, such as the track & trace system. Defining the components helps to create models and analysis simulation studies. In order to be able to model Physical Internet-based intralogistics systems, we consider it essential to define the components required for the system. The intralogistics conversion of each component is summarized in Fig. 2, which will be detailed in the following subsections. We review the previously identified components, covering which can be convertible to the intralogistics systems and which require change. Furthermore, in the case of intralogistics systems, we explore the applicability of current solutions and system elements, as well as the compatibility of the new technological achievements introduced by Industry 4.0.

3.1 *π-Container*

One of the most important components of a Physical Internet-based system is the π-container responsible for storing, moving, and transporting goods. It is very important that these units meet environmental requirements, one of the most significant parameters of which is sustainability. In order to achieve this goal, an environmentally friendly, flexible, reusable and robust design is essential. In addition, it must have an additional IT-based feature. This is related to the identification and the information carried by the unit and its accessibility. Another important feature is the traceability of π-containers at all times.

The various plastic crates, totes, boxes are still widely used in intralogistics systems. In reference [6] the a π-containers were extended into smart pallet. The

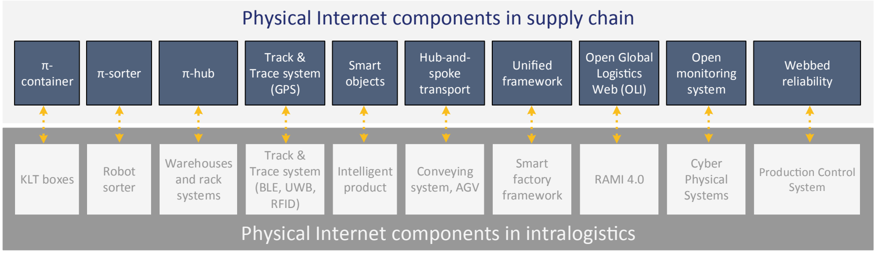

Physical Internet components in supply chain
π-container
π-sorter
π-hub
Track & Trace system (GPS)
Smart objects
Hub-and-spoke transport
Unified framework
Open Global Logistics Web (OLI)
Open monitoring system
Webbed reliability
KLT boxes
Robot sorter
Warehouses and rack systems
Track & Trace system (BLE, UWB, RFID)
Intelligent product
Conveying system, AGV
Smart factory framework
RAMI 4.0
Cyber Physical Systems
Production Control System
Physical Internet components in intralogistics

Fig. 2 Intralogistics conversion of the Physical Internet components

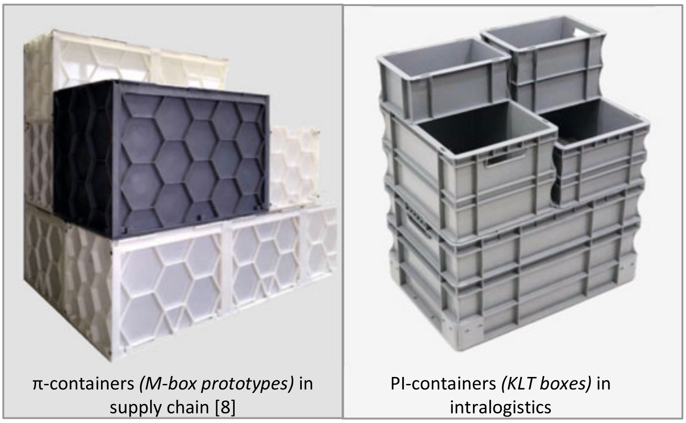

π-containers *(M-box prototypes)* in supply chain [8]
PI-containers *(KLT boxes)* in intralogistics

Fig. 3 Intralogistics conversion of π-container

most commonly used units in manufacturing are the KLT boxes. KLTs basically have standard sizes, which are combinations of the size of EUR1 pallet in order to combine together [7]. Based on a very similar idea, π-containers were defined based on the size of ISO containers [8].

The purpose of the KLT is to create a uniform handling of goods. In addition, by identifying and tracking the KLT, it is also possible to track the products or parts to be stored in them. In our opinion, this has already fundamentally created the conversion of a π-container that can be integrated into a Physical Internet-based system.

Of course, in order to be integrated into an intelligent system, it must be transformed into "smart" KLT box, an example of which can already be seen in the smart factories that appeared in connection with the Industry 4.0 concept. Figure 3 illustrates the π-container to be used in the supply chain and the intralogistics conversation of the component, the KLT boxes.

3.2 π-Hub

The next component of the future logistics network is the π-hub, which is an open and accessible transit center for all participants in the PI system [9]. These hubs represent the main opportunity for cooperation, where π-containers are joint to the appropriate vehicle, taking into account economic, social and environmental considerations. An important feature of organizational issue is that by creating compatibility, each unit can be associated with any resource. Simulation studies have shown that π-hubs created in PI have a positive effect on the performance of facilities: they reduce the waiting time of trucks and the waiting time of π-containers [10]. PI-Hubs reduce loading and unloading time by 90% and resource consumption by 87% [11]. There are two versions of π-hubs. One is a fixed physical center with infrastructure and the other is a virtual transfer point.

1. *Fixed π-hub*

In most models and analyzes, fixed and predefined centers are created in the Physical Internet system. These transit points include a number of functional elements such as parking bays, sorters or services [12]. In the following, we examine which types of warehouse related to production can correspond to the defined π-hub. We can define the following types of warehouses related to production:

- Raw material and spare parts warehouses
- Auxiliary material warehouses
- Semi-finished goods warehouses
- Finished goods warehouses.

In our opinion, of these types of warehouses, the semi-finished goods warehouse is clearly a center inside an intralogistics system that can be defined as a π-hub. The raw material and spare parts warehouses and auxiliary material warehouses can also be defined as a π-hub conversion, as we are faced with a node with multiple connections in terms of raw materials and empties. Since goods are received from the intralogistics side to the finished goods warehouse from only one direction, we do not consider it necessary to convert this into a π-hub.

It is worth examining the compatibility of the applicable storage methods with the Physical Internet system. In terms of storage, in our opinion, storage without a rack, drive-in racking, drive-through racking, mobile pallet racking and selective racking cannot be integrated into a PI-based system, as it does not provide the necessary flexibility and combinability and is very slow to operate. The channel-based racking (shuttle car system) is really recommended for integration into a PI-based intralogistics system. This provide the expected flexibility and fast service, and it is important that material handling can be automated. In the case of vertical or horizontal carousel storage methods, the inbound and outbound process are on the same side. This makes it less flexible than the previous storage mode, but we think it can be integrated into a PI-based system for lower intensity and less complex points.

1. *Virtual π-hub*

Virtual π-hubs are transfer points in the network whose location is not fixed and has no infrastructure elements. It is designed to complement fixed π-hub centers to provide additional meeting space for resources to achieve better utilization, shorter distances, and faster delivery at the system level. In the network, this cooperation can be applied, for example, between platoons formed by heavy-duty vehicles, where platoons can reconfigure themselves by changing vehicles, thus achieving a more ideal state [13].

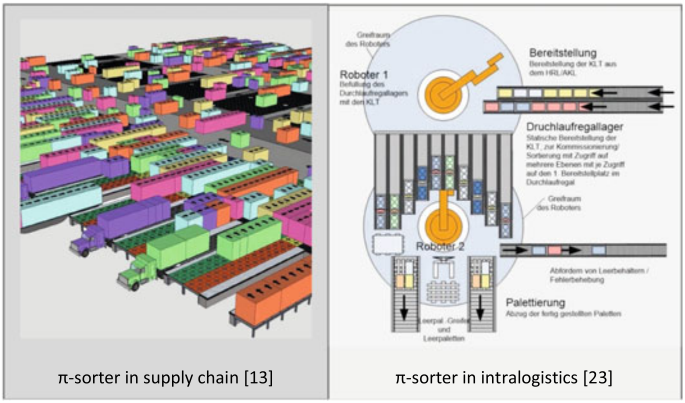
The virtual hub integrated in the supply chain has a fundamentally low applicability for intralogistics systems. This is because in intralogistics systems, a much shorter distance must be covered when examining the supply chain. For this reason, we need a minor modification to integrate the component. In intralogistics systems, a transit point without a storage system is common, where significant goods are delivered and send. These points can correspond to the virtual π-hub.

It is important to mention tandem AGV systems where AGVs transfer goods to each other in connection with the possibility of intralogistics conversion of virtual π-hubs. The required meeting place in the system can be variable and virtual, so there are no designated locations [14].

3.3 π-Sorter

The π-sorters are important parts of π-hubs that provide the sorting and assignment function. Its task is to receive π-containers from different points and classify them by assigning each to a specific truck in a potentially defined order [15]. The π-sorter may include a network of conveyors and/or other π-sorters. In several simulation studies, the grid sorter modular conveyor and sorting system was integrated into π-hubs for efficient and space-saving classification and movement of goods.

There is no doubt that sorting tasks come with a high intensity in intralogistics systems as well. Another important aspect for integration into a Physical Internet-based system is flexibility and efficiency that automation can help. Based on this, the intralogistics conversion of π-sorters can be clearly matched by sorting and palletizing robots [16], even by combining the previously mentioned conveyor systems, which is illustrated in Fig. 4.


Greifraum des Roboters
Roboter 1
Befüllung des Durchlaufregallagers mit den KLT
Bereitstellung
Bereitstellung der KLT aus dem HRL/AKL
Druchlaufregallager
Statische Bereitstellung der KLT, zur Kommissionierung/ Sortierung mit Zugriff auf mehrere Ebenen mit je Zugriff auf den 1. Bereitstellplatz im Durchlaufregal
Greifraum des Roboters
Roboter 2
Abfordern von Leerbehältern / Fehlerbehebung
Palettierung
Abzug der fertig gestellten Paletten
Leerpal. Greifer und Leerpaletten

π-sorter in supply chain [13]

π-sorter in intralogistics [23]

Fig. 4 Intralogistics conversion of π-sorter

3.4 Smart Objects

The Physical Internet takes advantage of embedding smart objects. Each π-container comes with a smart tag that represents a unique identifier worldwide, similar to the MAC identifier known from the digital internet system. This smart tag ensures identification, integrity, control and traceability. Smart objects also provide the ability to automate handling, storage and control. In addition, smart objects communicate and interact with each other and are able to make independent decisions and self-control [15].

Intralogistics conversion of smart objects flowing in a network is an intelligent product emerging with the advancement of technology. An intelligent product is a physical and virtual based representation of a product. The connection between the two is made using the label and the reader. Intelligent products constantly monitor their condition and environment, based on which they can respond by considering operating constraints. They are also able to memorize, communicate, and trigger events, or notify users when a problem is detected [17]. Within the PI-based shop floors, smart manufacturing objects (SMOs) appear, which use Internet of Things (IoT) technology to create intelligent manufacturing [6].

3.5 Hub-and-Spoke Transport

Current transport is characterized by a mixed use of point-to-point and hub-and-spoke transport. This is mainly due to the lack of cooperation, with which independent, direct deliveries result in a high number of empty flights and further pollution. To improve this, in a Physical Internet-based network, deliveries would operate as part of an interconnected network with a hub-and-spoke solution. This defines the first non-physical component element for Physical Internet-based systems. The key is to create collaboration and flexibility. Many nodes (π-hubs) available to PI participants allow for a smooth flow of π-containers from hub to hub, independent of the container owner [18].

The intralogistics systems can be defined in the following grouping according to the material handling method used [19]:

- Direct transportation (logistics-taxi): transport takes place only between the point of sending and receiving, which can be manually or with equipment, like AGV or different type of forklifts.
- Milk-run system (logistics-bus): a cyclical transport system that delivers different goods, raw materials and packages to different points based on a route and schedule fixed during the run.
- Conveying system (logistics-train): transport in one direction provided by continuous material handling equipment between the sending and receiving points.

From the listed methodologies, we examined which applications can be integrated into the system defined by the Physical Internet. In recent years, there has been a growing interest in decentralized control systems such as the CTS (cellular transport system). The CTS system is based on material handling entities, which can also correspond to the logistics-taxi or logistics-train concept [20]. The decentralized system provides the flexibility, performance, tracking and communication needed to integrate into a Physical Internet-based system. Therefore, in our opinion, the logistics-bus concept cannot be equated with the intralogistics conversion of the PI component.

3.6 Track and Trace System

An essential component for achieving an open and interconnected system is track and trace system (T&T) that ensures the smooth flow of all objects in the network. This is based on identifying and tracking the past and present locations of objects in the network (π-containers, trucks), for which each container must be uniquely labeled. Continuous outdoor tracking is enabled by GPS (Global Positioning System) or mobile tracking based on location via a mobile connection. Outdoor tracking with Wi-Fi is not possible, so in this case only 4G or 5G telecommunication technology can be used [21].

The use of the T&T system is not new in intralogistics systems either, as it is one of the foundations of the smart factories and warehouses introduced by Industry 4.0. The most common indoor tracking methods are barcode and RFID. In the case of fixed scanner units, information transmission can be implemented via Wi-fi, which is a cost-effective solution based on the current state of the art. The most popular wireless communication solutions for real-time positioning systems (RTLS) are Bluetooth and UWB [21]. In this case, the use of 4G is uncommon. With 5G technology, however, a reliable and secure network can be built, ensuring high connectivity and data density. The intralogistics conversion of the T&T system is shown in Fig. 5.

3.7 Unified Framework

The components of the Physical Internet include the creation of a unified framework that ensures the global operation of the system. It should be based on the same conceptual framework, regardless of the size of the network involved. The system can be identified with the style of multilevel Russian dolls, where the networks operate according to the same protocol and standard [15]. PI uses the same conceptual framework for the facility level, the regional, the city level and even for the global level.

In the case of intralogistics conversion, the development of the framework can also be considered a necessary component, as it is essential for communication and

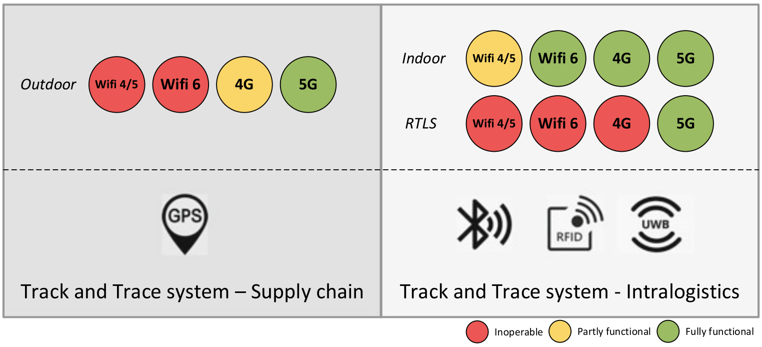


Fig. 5 Intralogistics conversion of Track & Trace system

standard operation. The layer-based framework used in PI also appears here, only slightly differently. Each layer starts from the sensing required in the physical world, and then the decision-making and control layer can be defined through the interface layer in order to achieve better job cooperation and collaboration between the physical and cyber worlds [22].

3.8 Open Global Logistics Web

The logistics web defined by the Physical Internet must be an open, global, efficient, and sustainable system. The Open System Interconnection (OSI) reference model adopted by the International Organization for Standardization (ISO) is used to introduce the PI-based Open Logistics Interconnection (OLI) reference model [11].

To define intralogistics conversion, we use the Reference Architectural Model Industry 4.0 (RAMI 4.0) published by Industry 4.0. The model provides a uniform interpretation for standards and use cases [23].

3.9 Open Monitoring System

The logistics network of the future must be equipped with an open monitoring system for its decentralized and autonomous management. This allows us to provide monitoring, communication and feedback. Its intralogistics conversion is Cyber-physical systems (CPS), which are systems of cooperating computing units. The units are intensively connected with the processes of the physical world, with which at the same time its virtually appearing data access and processing can be realized [23].

3.10 Webbed Reliability

The last component of the Physical Internet is to provide security. The entire PI network must guarantee its own reliability. Ensuring the security of containers and shipments within the protocol and structure. A system of multiplied nodes in a world-class network must provide robustness and resilience.

Due to the layered system defined as the intralogistics conversion of the "Unified framework", reliability is one of the challenges that needs to be addressed. In intralogistics systems, this function must be guaranteed by the control system.

4 Simulation Model of an Intralogistics PI Network

There are serious doubts questioning applicability of the PI concept in factory logistics because of the vast differences to supply chain networks. The core differences are complexity and delivery times. Former comes from the fact that in supply chain networks a larger dimension of suppliers and customers are networked in a unified system, with different production schedules. However, inside a company there is a far less complex production plan which may not necessitate complex material flows. Regarding delivery times, these are inside in a production facility, it takes usually some minutes, so the necessity of additional transfer in π-hubs are questioned.

Advantages of a PI system in intralogistics are the separately established PI structure which carries out the transports more effectively than the separate material transport services for each source-drain pairs. Having this conventional transport network, the question is always the frequency of transports. High frequency transports would result in low inter production stocks and shorter throughput times, however causes application of low capacity or not fully utilized material handling machines, which means higher specific material handling costs. We expect that operating in a PI network will increase the material handling machines utilization, which means at the same time lower traffic intensity and fewer necessary human and machine resources.

As the advantage of the PI system over the conventional is not evident detailed simulation studies are necessary. For the comparison of the conventional and PI systems, we defined performance level for the measurement of ineffectiveness as follows.

$$Ineffectiveness = \sum_{k=1}^{n} \left(C_{free,k} S_k \right) \quad (1)$$

where

- $C_{free,k}$ is the free capacity in loading units during the kth transport task, and
- S_k is the length of the kth transport route.

Matching the PI material handling network to the factory structure varies strongly case for case, therefore a single example has been selected for explanation.

Figure 6 depicts an imagined factory with two assembly lines where different products are manufactured. There are four types of routes for the components. Directly supplied materials are transported from the raw materials warehouse to the assembly lines using separate material handling activity (1). The second type of components are first transported from the raw warehouse (2), machined, and transported to the semi-finished goods warehouse (4) before transported to Assembly 1 and 2 (6). The third type of components are painted parts transported via (3), (5) and (6). The fourth part type is of components first machined and afterwards painted, before transported to the assembly via (6). All transport relations are bidirectional: containers (e.g. KLTs) are streaming in one direction with full parts, and empty containers are transported back, which is common in intralogistics. Having a PI network the main difference is that different component types (e.g. supplied and machined parts) can be transported on the same material handling machine at the same time ensuring short throughput times and high utilization.

For the comparison of the conventional and PI systems regarding the ineffectiveness we propose use of simulation models. During our research a simulation

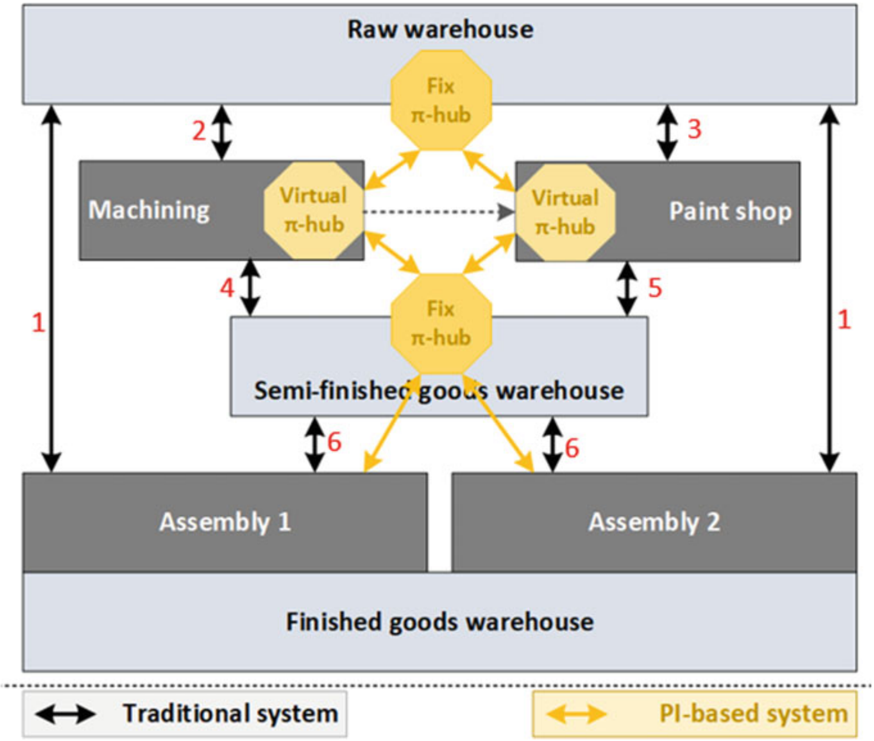


Fig. 6 Example of a PI network's implementation in factory environment

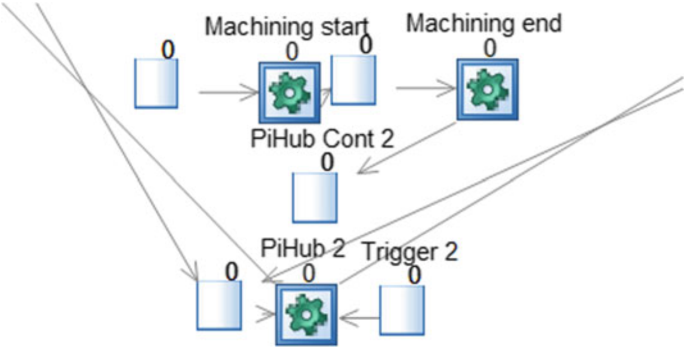


Fig. 7 Simulation model's detail of a PI based intralogistics system

model has been developed which will be a good basis for the analyses of the coming research activity. The model's development has been carried out using Simul8, which is a DES simulation software. We would like to point out that the simulation model has a relatively simple structure, concentrating on the material flows. Figure 7 shows the part, which models the machining part with a possibility of setting an average delay as machining time for all parts. Material handling machines (forklifts, AGVs) are approaching into the components "PiHub". Materials stored in the hub and transported materials are in "PiHub Cont".

The model's operating logic is simple. Each hub checks in uniform time intervals amount of transport needs to other hubs and clients. Next available vehicle is dispatched for the station with the largest demand.

5 Summary

The concept presented in this paper is a relatively new approach with only some related references for intralogistics environment. As no direct comparison studies have been found to conventional systems, next step of our research activity concentrates on the use of the simulation model for detailed analysis, in order to specify for which factory structure is it beneficial.

References

1. H. Treiblmaier, K. Mirkovski, P.B. Lowry, Z.G. Zacharia, The Physical Internet as a new supply chain paradigm: a systematic literature review and a comprehensive framework. Int. J. Logist. Manage. 1–49 (2020)
2. E. Karakaya, Y. Kayikci, F. Öztürk, Implementation oh Physical Internet-based intelligent intralogistics systems: an agent-based simulation study, in *International Symposium for*

Production Research, 2017, pp. 379–392

3. B. Montreuil, R.D. Meller, E. Ballot, Physical Internet Foundations. IFAC Proc. **45**(6), 26–30 (2012)
4. H. Treiblmaier, K. Mirkovski, P.B. Lowry, Conceptualizing the Physical Internet: literature review, implications and directions for future research, in *11th CSCMP Annual European Research Seminar*, Vienna, Austria, 2016
5. E. Puskás, G. Bohács, Physical Internet—a novel application area for industry 4.0. IJEMS (Int. J. Eng. Manage. Sci.) **4**(1), 152–161 (2019)
6. R.Y. Zhong, C. Xu, C. Chen, G.Q. Huang, Big data analytics for Physical Internet-based intelligent manufacturing shop floors. Int. J. Prod. Res. 2610–2621 (2015)
7. B. Montreuil, E. Ballot, W. Tremblay, Modular design of Physical Internet transport, handling and packaging containers, in *Progress in Material Handling Research: International Material Handling Research Colloquium*, 2014
8. C. Landschützer, F. Ehrentraut, D. Jodin, Container for the Physical Internet: requirements and engineering design related to FMCH logistics. Logist. Res. **8**(1) (2015)
9. S. Pan, M. Nigrelli, E. Ballot, R. Sarraj, Y. Yang, Perspectives of inventory control models in the Physical Internet: a simulation study. Comput. Ind. Eng. 84 (2014)
10. T. Chargui, A. Bekrar, M. Reghioui, D. Trentesaux, Simulation for PI-hub cross-docking robustness, in *Service Orientation in Holonic and Multi-Agent Manufacturing*, 2018, pp. 317–328
11. S. Gontara, O. Korbaa and A. Boufaied, "Routing the Pi-Containers in the Physical Internet using the PI-BGP Protocol", *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, 2018
12. B. Montreuil, R.D. Meller, C. Thivierge, Z. Montreuil, Functional design of Physical Internet facilities: a road-based crossdocking hub, in *12th IMHRC Proceedings* (Gardanne, France—2012), 2012, p. 42
13. E. Puskás, G. Bohács, Concepting freight holding problems for platoons in Physical Internet systems. Acta Logist. Int. Sci. J. Logist. **6**(1), 19–27 (2017)
14. B. Rahimikelarijani, H. Fazlollahtabar, S. Nayeri, Multi-objective multi-load tandem autonomous guided vehicle for robust workload balance and material handling optimization. SN Appl. Sci. **2**, 1200 (2020)
15. B. Montreuil, Toward a Physical Internet: meeting the global logistics sustainability grand challenge. Logist. Res. **3**, 71–87 (2011)
16. U.A. Lammer, Funktionsvereinigung in der Lagertechnik", *Lehrstuhl für Fördertechnik Materialfluß Logistik (fml) TU München*, 2017
17. G.G. Meyer, K. Främling, J. Holmström, Intelligent products: a survey. Comput. Ind. **60**(3), 137–148 (2009)
18. T.G. Crainic, B. Montreuil, Physical Internet enabled hyperconnected city logistics. Transp. Res. Procedia **12**, 383–398 (2016)
19. W.A. Günthner, J. Durchholz, E. Klenk, J. Boppert, *Schlanke Logistikprozesse: Handbuch für den Planer* (Springer, Berlin, 2013)
20. E. Karakaya, Y. Kayikci, F. Öztürk, *Implementation of Physical Internet-based intelligent intralogistics systems: an agent-based simulation study* (International Symposium for Production Research, Vienna, 2017)
21. J. Menning, L. Hajek, P. Münder, 5G in Production, *Umlaut White Paper*, Germany, 2019
22. R.Y. Zhong, Design and development of a Physical Internet-enabled smart factory for discrete manufacturing, in *Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management* (2015)
23. T. Lins, R.A.R. Oliveira, Cyber-physical production systems retrofitting in context to industry 4.0. Comput. Ind. Eng. 139 (2020)

A Data-Driven Framework for Exploring the Spatial Distribution of Industries

Huifeng Sun

Abstract Having a good understanding of the spatial distribution of industries, e.g., 5G, IT and New Energy, is of high importance for each country. This work thus proposes a general data-driven framework to explore and demonstrate such a distribution. First, we integrate data from different sources and build a big data store for analyzing industries. Then we develop a industry data query processing module and an industry spatial distribution analytic module based on the built data store to provide efficient queries (e.g., spatial query, keyword query and hybrid query) and intelligent data analysis (e.g., heterogeneous data fusion, industry clustering analysis, and company clustering analysis). In addition, we also develop a visualization interface to illustrate the querying and analysis results. As validated by the experiments over a real dataset, the proposed framework can well capture the spatial distribution of various industries and gives a new view of the development of industries in certain region or country.

Keywords Industry and company · Spatial distribution of industry · Spatial clustering · Visualization

1 Introduction

The rapid development of various industries plays an indispensable role in national economy. Having a good understanding of the spatial distribution of industries can help to identify the bottlenecks in supply chain and support decision-making for economic policy and plans [1].

Existing studies on analyzing the distribution of industries mainly focus on specific administration regions and industrial parks. However, with the development of modern transportation and logistics, the boundary between administration regions is becoming more and more vague for various industries. For example, with the

H. Sun (✉)
Beijing Academy Of Artificial Intelligence BAAI, Beijing, China
e-mail: hfsun@baai.ac.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_68

help of e-commerce, one company could sell its products across the whole country, even all over the world, and competes with other competitors. As for the government, regional integration has been an important part in making the economic plans. Therefore, having a view of global distribution of industry is of high necessity and practical significance.

In this work, we propose a general data-driven framework to analyze the spatial distribution of industry. Under this framework, industry data from multiple sources will be collected and integrated to build a big industry data store, efficient query processing module will be developed to support quick data access, and spatial distribution analysis for industry will be conducted from different aspects. In addition, a visualization interface will be developed to provide a better understanding of the spatial distribution of industry.

The rest of this work is organized as follows. Section 2 briefly reviews the related work. Section 3 overviews the proposed framework and describes the dataset that will be used as an example. Section 4 elaborates the technical details for each module in the proposed framework and shows the analysis results. Finally, Sect. 5 concludes the work and highlight the future directions.

2 Related Work

This section briefly reviews the related work on analyzing the distribution of industries. Existing methods mainly focus on the distribution of specific industries, e.g., food, block chain, tyre, and drug.

Concretely, Zhou et al. [2] proposed an improved k-meands algorithm with penalty to analyze the distribution of restaurant industry. Bacchetti et al. [3] conducted a case study to study the way to improve the distribution plan process in the food and beverage industry. Liang et al. [4] developed a smart inventory management system for food processing and distribution industry. All these studies are targeted at food industry.

In addition, Friedlmaier et al. [5] studied the regional distribution of Block Chain industry; Mahfouz et al. [6] proposed an assessment framework for the tyre distribution industry; and Guo et al. [7] evaluated the evolution of Chinese drug distribution industry.

Obviously, all the existing studies above only focus on some specific industries and it is usually difficult to apply the proposed methods to other industries. Therefore, in this work, we aim to propose a general data-driven framework to analyze the spatial distribution of various industries, which makes it different from the existing studies.

3 Framework Overview and Dataset

3.1 Framework Overview

Figure 1 illustrates the proposed framework for exploring the spatial distribution of industries. This framework is consisted of multiple function modules, including multi-source data storage and index module, industry data query processing module, industry distribution analytic module and visualization interface. Concretely, we first develop a multi-source data storage and indexing module to collect spatial data, region data, company data, etc. Second, we develop a industry data query processing module to support efficient processing for spatial query, keyword query and hybrid query. Third, we develop a big data analytic module which can achieve heterogeneous data fusion, industry clustering analysis, and company clustering analysis. Finally, a visualization interface will be developed to provide visualization for spatial maps, query results, and analytic results.

The technical details for these modules will be elaborated in Sect. 4.

3.2 Dataset

In this work, we use the listed companies in China as an example to present our idea for exploring the spatial distribution of industries.

In the dataset, there are 3,273 companies listed at the Chinese stock market. Figure 2 shows the number of listed companies in each province. According to the

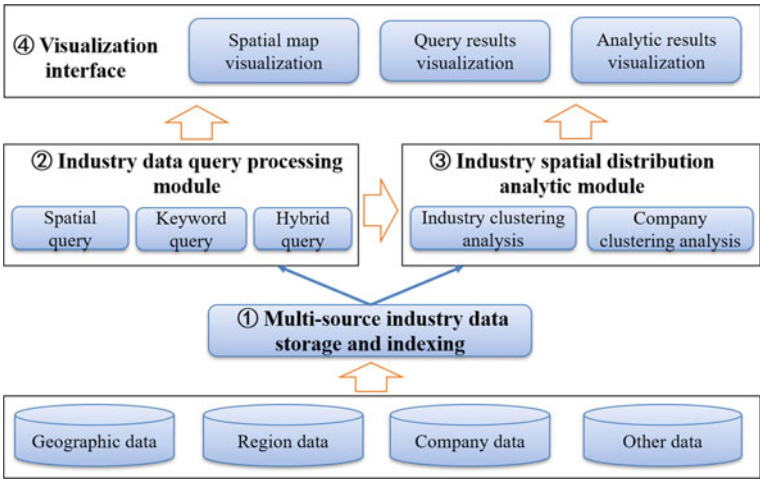

④ Visualization interface
Spatial map visualization
Query results visualization
Analytic results visualization
② Industry data query processing module
Spatial query
Keyword query
Hybrid query
③ Industry spatial distribution analytic module
Industry clustering analysis
Company clustering analysis
① Multi-source industry data storage and indexing
Geographic data
Region data
Company data
Other data

Fig. 1 The data-driven framework for exploring the spatial distribution of industries and entities

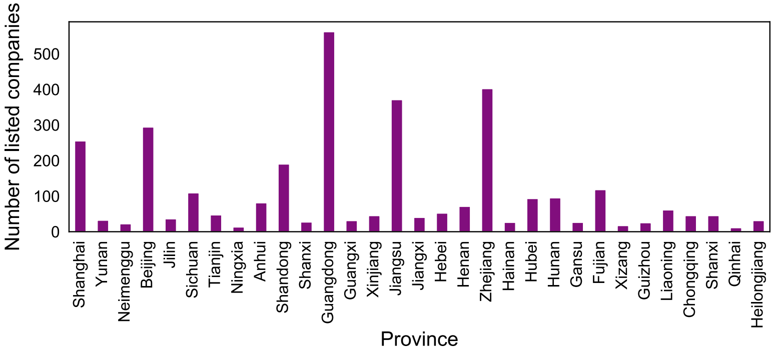


Fig. 2 The number of listed companies in each province

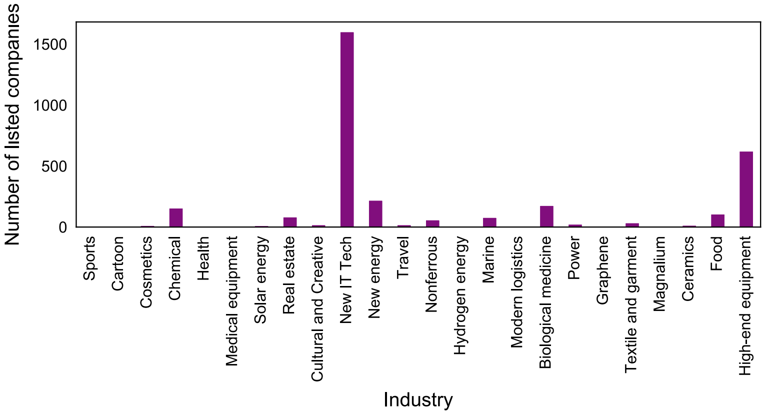


Fig. 3 The number of listed companies in each industry

figure, Guangdong, Jiangshu, Zhejiang, Beijing and Shanghai have the top-5 biggest number of listed companies.

We have 28 major industries in this dataset. Figure 3 illustrates the distribution of listed companies over different industries. Obviously, New IT Tech has the biggest number of listed companies and High-end Equipment follows.

4 Detailed Techniques

4.1 Multi-source Industry Data Storage and Indexing

To support efficient industry big data querying and analysis, we design a multi-source industry data storage and indexing module. Figure 4 illustrates the architecture of the

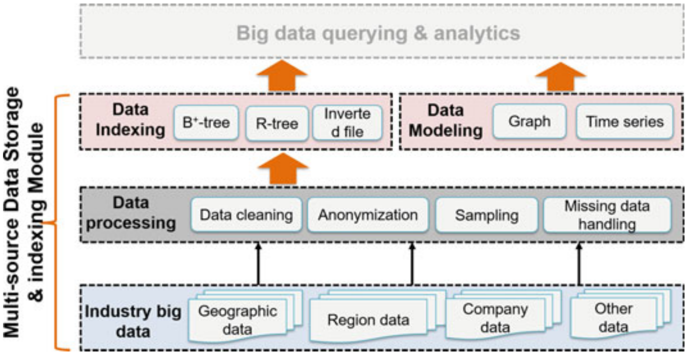


Fig. 4 The architecture of multi-source industry data storage and indexing module

designed module which consists of multiple layers. The bottom layer is the data from different sources, such as geographic data, region data, and company data. Since the data could be of low quality and has certain privacy concerns, we develop the data processing layer with a series of functions, e.g., data cleaning, anonymization, sampling, and missing data handling.

To support efficient query processing for industry big data, we create different indices over the data according to the data characteristics and types of query. For traditional search query over certain attributes, we build the widely used B$^+$-tree over these attributes; for spatial query, we build an R-tree to index the spatial attributes; for keyword query, we build an Inverted File index to speed-up the query.

In addition, before analyzing the industry big data, we need to represent the data with certain data structures. For example, to model the relation between different industries, we can build a graph over all industries, use node to represent industry, and use link between nodes to indicate their relation. The historical records of each industry can be represented as a time series to capture the temporal dependency.

4.2 Industry Data Query Processing Module

Analyzing the spatial distribution of industries requires efficient query processing to facilitate the data acquisition. Therefore, we design an industry data query processing module to provide multiple types of queries for industry data as below.

- **Spatial query**: Spatial query aims to search industry data according to the spatial attributes. For example, we can use a range query to search all the 5G companies in a given spatial region. To achieve efficient processing for spatial query, we design query processing algorithms based on the R-tree indexing and utilize pruning mechanisms to reduce the search space.
- **Keyword query**: Given a set of keywords, keyword query aims to find out all the records that contain the keywords or are highly relevant to the keywords.

For keyword matching query (i.e., containing query keywords), we design a query processing algorithm based on Inverted File index which first identifies the industry records containing each query keyword, and then integrates these records together to generate the final results. As for relevant query, we use TF-IDF model [8] to quantify the text relevance between query keywords and the potential query results. The industry records with the highest relevance are returned as the final query results.

- **Hybrid query**: Hybrid query searches industry records based on both spatial proximity and keywords, which makes it more challenging to achieve efficient query processing. To address this issue, we build a hybrid index IR-tree [9] which is a combination of Inverted File and R-tree. With IR-tree, we further design a hybrid query processing algorithm to achieve spatial and textual pruning at the same time.

4.3 Industry Spatial Distribution Analytic Module

The spatial distribution of industries has two aspects, i.e., the spatial distribution of companies for each industry and the spatial distribution of different industries. Therefore, we will elaborate the analysis from these two aspects as below.

(a) Spatial distribution of companies for each industry
Without loss of generality, we assume that each company c is represented as $c = (id, name, t, lon, lat)$ where id is the identifier of the company, $name$ is the name of the company, t is the industry type that this company belongs to, lon is the longitude, and lat is the latitude. For example, Bank of China (BOC) belongs to new generation of IT technology, and its longitude and latitude are 114.107636E and 22.540466N, respectively.

Each industry I contains a set of companies of different sizes and we aim to uncover the spatial distribution of these companies.

For spatial distribution of companies in each industry, the most critical issue is to identify the clustering patterns among all companies. To address this issue, we introduce unsupervised clustering methods to group companies according to their longitude and latitude. Considering that the number of companies could be very large, we utilize Mini Batch K-means [10] to conduct clustering since it is more efficient than other clustering methods, especially for large datasets.

One challenge for clustering companies in each industry is to set an appropriate value for the number of clusters, i.e., m. To overcome this challenge, we evaluate the inertia when using different numbers of clusters. Here, the inertia can quantify the cohesion of the produced clusters and usually the smaller the better.

For example, Fig. 5 shows the inertia for different numbers of clusters for the listed companies in New IT Tech industry which ranks the top-1 industry. According to the figure, the inertia decreases when the number of clusters increase and keeps almost the same for 25 or more clusters. Therefore, we set m to 25 in this work.

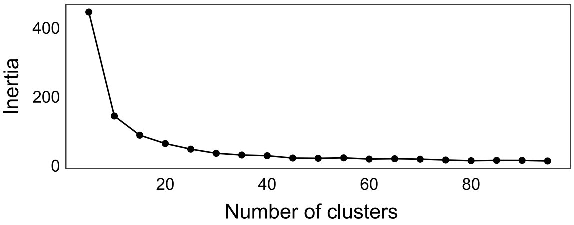


Fig. 5 The inertia for different numbers of clusters in Mini Batch K-Means

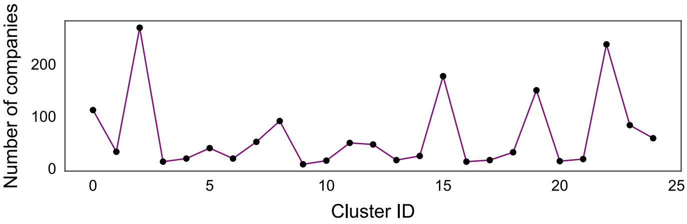


Fig. 6 The clustering results for companies in New IT Tech industry when setting $m = 25$

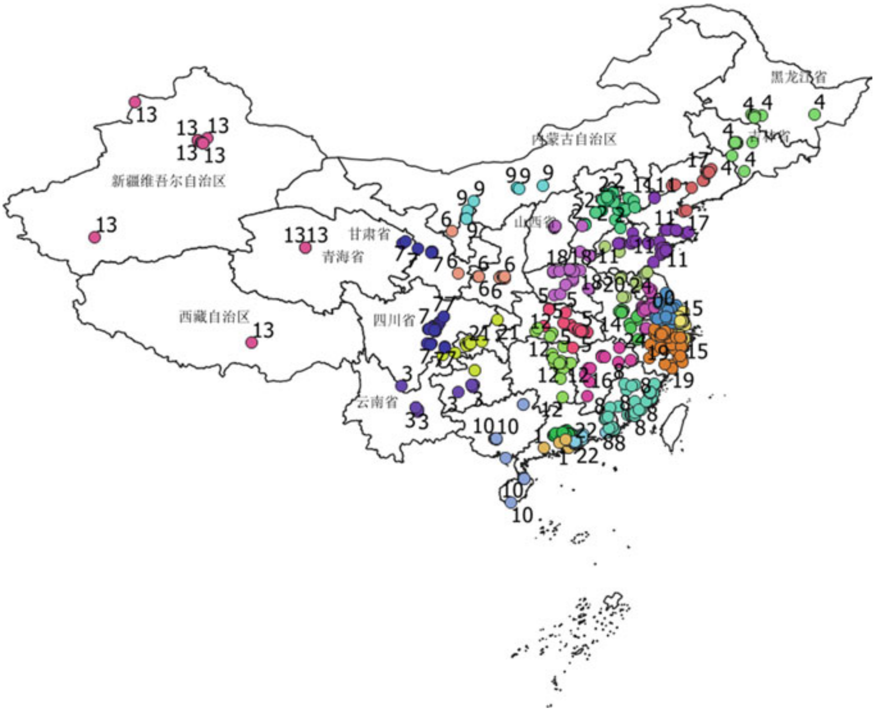


Fig. 7 The visualization of clustering results for companies in New IT Industry when setting $m = 25$

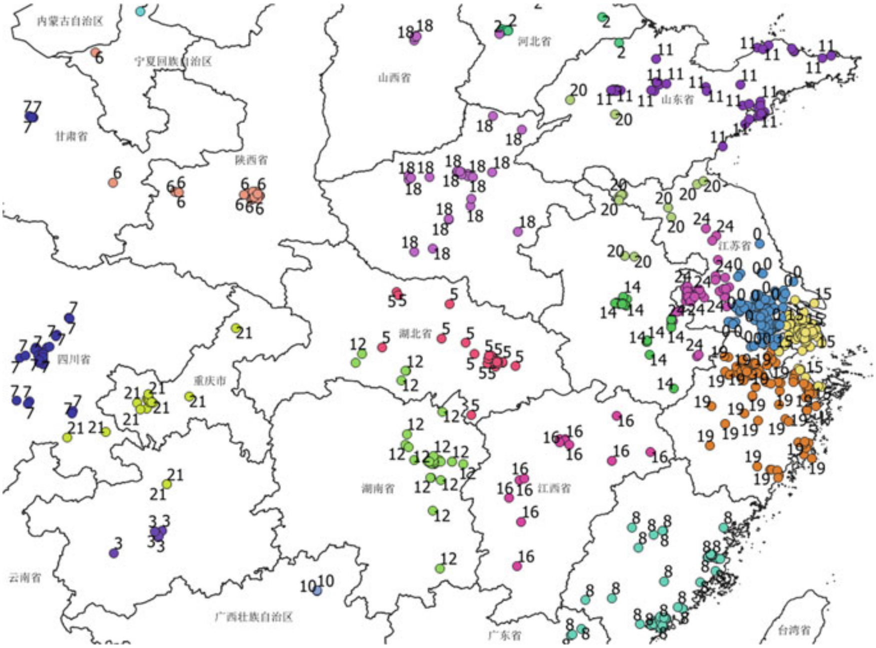


Fig. 8 The visualization of clustering results for companies in New IT Industry around the Yangtze River Delta

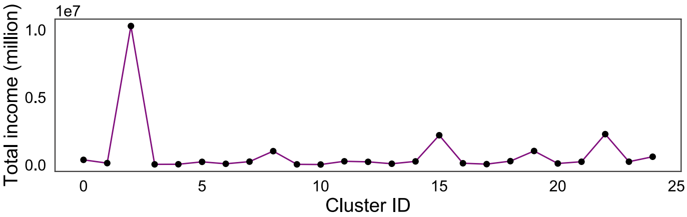


Fig. 9 The total income for each cluster of New IT Industry company in 2019

Figure 6 illustrates the number of companies in each cluster when setting the number of clusters to 25. As suggested by the figure, some clusters contain much more companies than other, which is because some well developed regions have more New IT companies than the less developed regions.

Figure 7 visualizes the generated clusters over map, where the number is the id of the corresponding cluster. Obviously, most of the New IT companies are located in East China, especially around the Yangtze River Delta. Figure 8 shows the clusters around the Yangtze River Delta.

Figures 9 and 10 present the total income and average income for each cluster of New IT Industry company in 2019. sThis cluster with the highest total income and

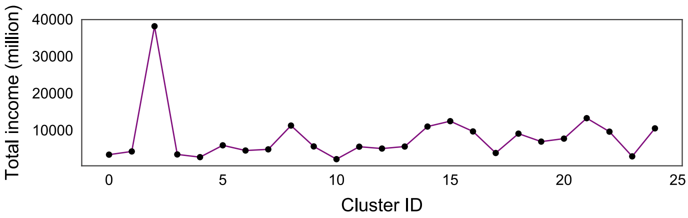


Fig. 10 The average income for each cluster of New IT Industry company in 2019

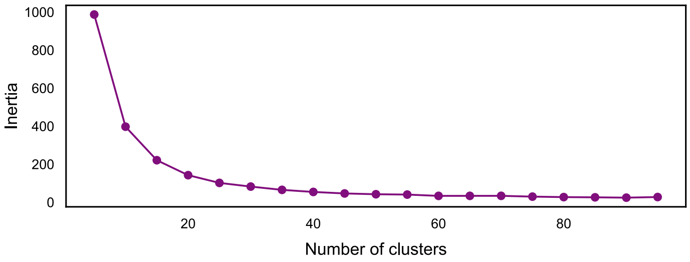


Fig. 11 The inertia for different numbers of clusters using Mini Batch K-Means for all listed companies at Chinese Stock Market

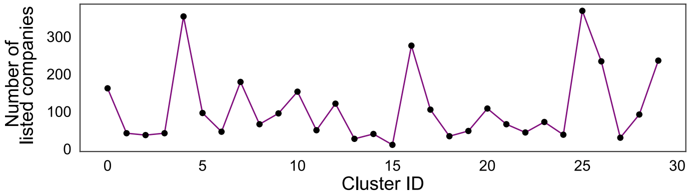


Fig. 12 The clustering results for all the listed companies when setting $m = 30$

average come is cluster 2 which corresponds the cluster around Beijing as illustrated in Fig. 7.

(b) Spatial distribution of different industries

Similar to the spatial distribution of companies for each industry, we also apply Mini Batch K-means to cluster all companies and analyze the industries in each cluster. According to Fig. 11, we set the number of clusters to 30.

Figure 12 illustrates the number of companies in each cluster and Fig. 13 visualizes the spatial distribution of those clusters around the Yangtze River Delta.

Particularly, we take the two clusters (cf. Figure 14) around Shanghai and Jiangsu as examples to analyze the distribution of industries.

Figures 15 and 16 illustrate the distributions of industries in the clusters around Shanghai and Jiangsu. As suggested by the figures, for both clusters, New IT industry and High-end Equipment industry rank the top-2 industries.

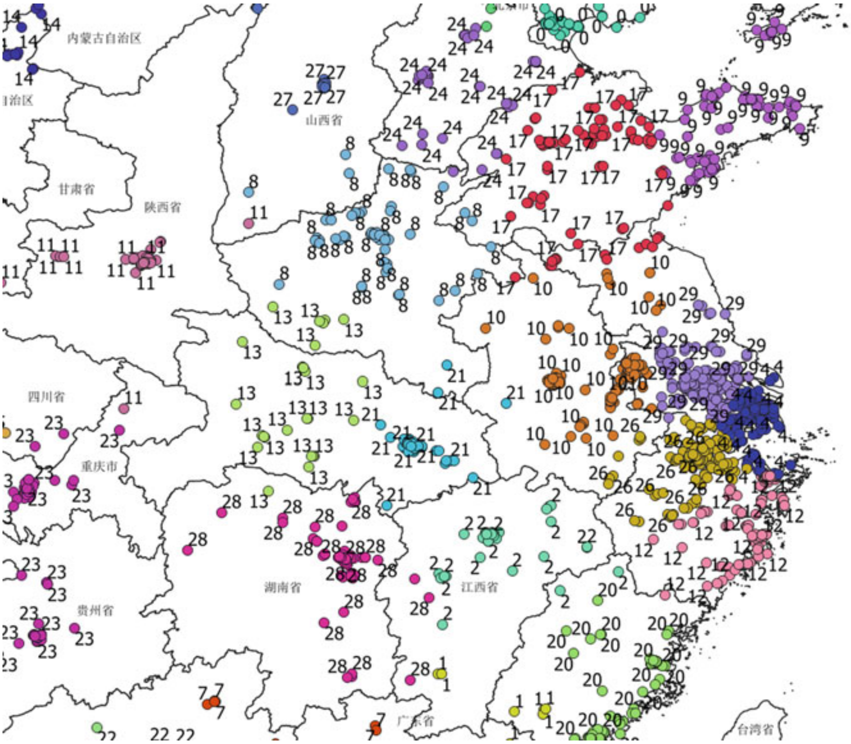


Fig. 13 The visualization of clustering results for the listed companies around the Yangtze River Delta

4.4 *Visualization Interface*

To achieve a better understanding of the analysis results, we develop a visualization interface to visualize the spatial distribution of companies and industries based Baidu Map. In the visualization interface, we design different layers, e.g., point distribution layer and heatmap distribution layer, to provide visualization from different views. For example, Fig. 17 illustrates the point distribution layer for all the listed companies.

5 Conclusion

In this work, we proposed a general data-driven framework for exploring the spatial distribution of industries. The proposed framework can fuse data from different sources to provide us insights in the development of various industries and benefit the optimization of supply chain and relocation of industries. In the future, we will further

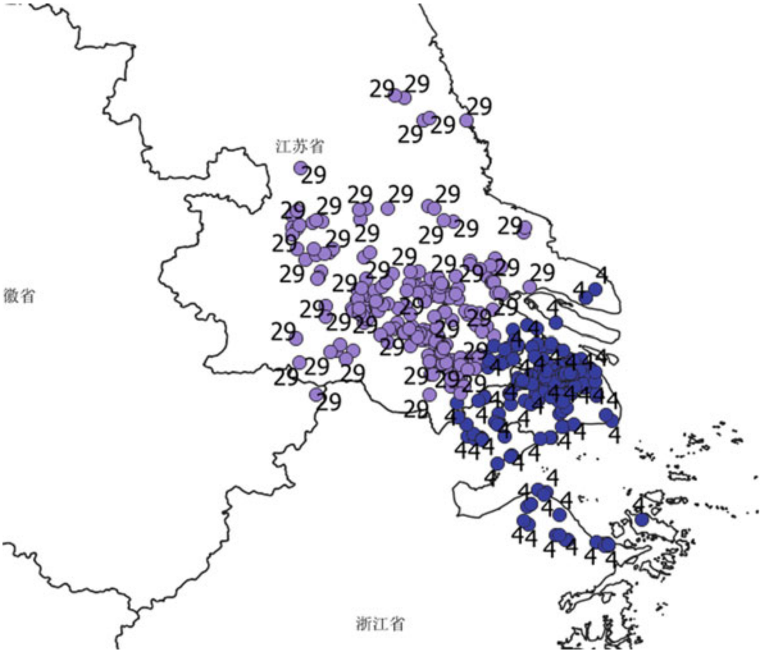
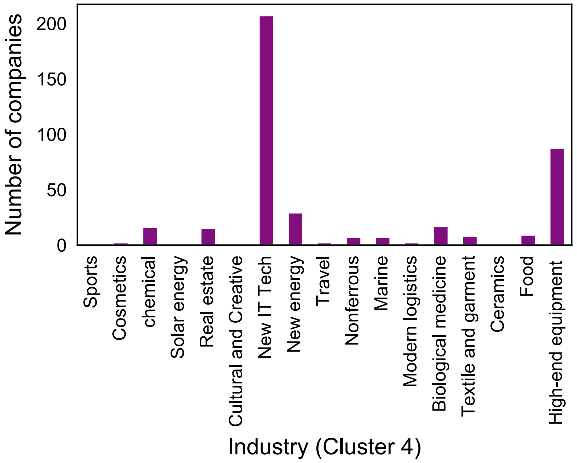


Fig. 14 The two clusters of listed companies around Shanghai and Jiangsu

Fig. 15 The distribution of industries in the cluster around Shanghai



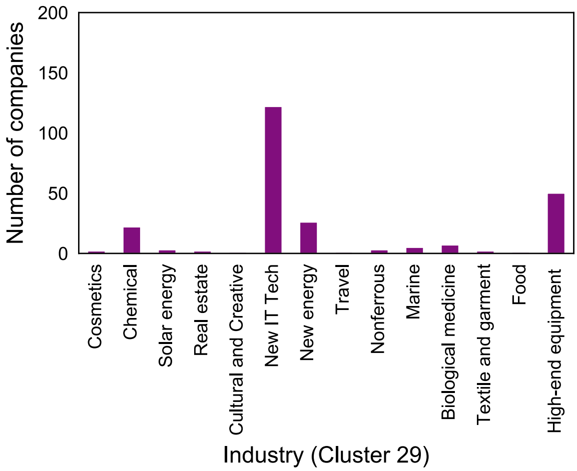


Fig. 16 The distribution of industries in the cluster around Jiangsu

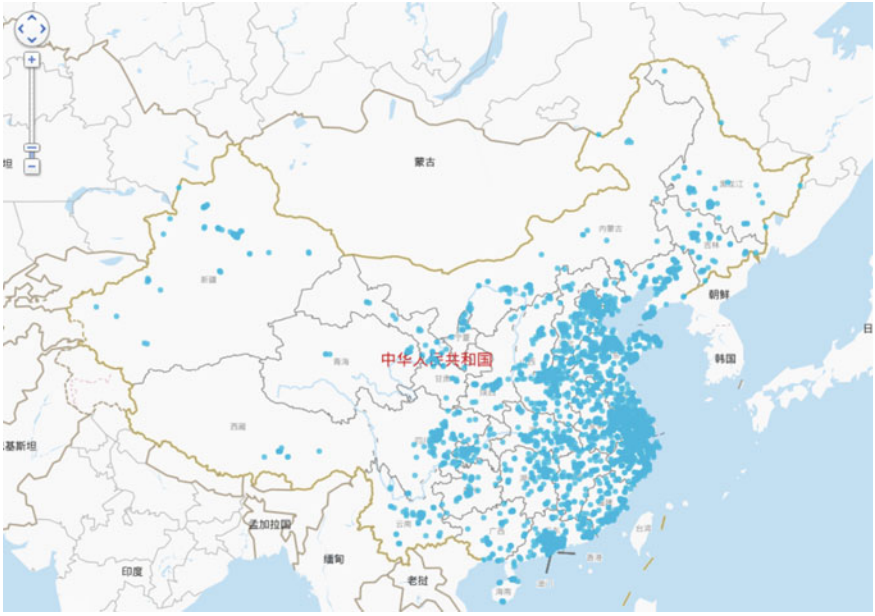


Fig. 17 The point distribution layer for all the listed companies

uncover the correlations between industries from different regions and identify the bottlenecks of coordinated development of regions.

References

1. B. Bilgen, Hans Otto Gïnther: Integrated production and distribution planning in the fast moving consumer goods industry: a block planning application. OR Spectr. **32**(4), 927–955 (2010)
2. Y. Zhou, R. Xie, T. Zhang, J. Holguin-Veras, Joint distribution center location problem for restaurant industry based on improved K-means algorithm with penalty. IEEE Access **8**, 37746–37755 (2020)
3. A. Bacchetti, M. Zanardini, Improving the distribution planning process in the food & beverage industry: an empirical case study, in *ECMS*, pp. 431–440 (2014)
4. C.-C. Liang, *Smart Inventory Management System of Food Processing and Distribution Industry*. ITQM, pp. 373–378 (2013)
5. M. Friedlmaier, A. Tumasjan, I.M. Welpe, *Disrupting Industries with Blockchain: The Industry, Venture Capital Funding, and Regional Distribution of Blockchain Ventures*. HICSS, pp. 1–10 (2018)
6. A. Mahfouz, A. Arisha, An integrated lean assessment framework for tyre distribution industry, in textitWinter Simulation Conference, pp. 3196–3197 (2015)
7. X. Guo, K. Reimers, B. Xie, M. Li, Network relations and boundary spanning: understanding the evolution of e-ordering in the Chinese drug distribution industry. JIT **29**(3), 223–236 (2014)
8. J. Zobel, A. Moffat, Inverted files for text search engines. ACM Comput. Surv. **38**(2), 6 (2006)
9. Z. Li, K.C.K. Lee, B. Zheng, W.-C. Lee, D.L. Lee, X. Wang, IR-tree: an efficient index for geographic document search. IEEE Trans. Knowl. Data Eng. **23**(4), 585–599 (2011)
10. D. Sculley, Web-scale k-means clustering, in textitProceedings of the 19th International Conference on World Wide Web, WWW, pp. 1177–1178 (2010)

The New Data-Driven Newsvendor Problem with Service Level Constraint

Yuqi Ye and Xufeng Yang

Abstract In today's data-rich world, decision makers can employ not only demand observations but also external explanatory variables (i.e. features) to solve the newsvendor problem without traditional demand distribution assumption, which has been drawing increasing attention and has derived so-called new data-driven approaches. Still in its infancy, this paper proposes an improved new data-driven method based on Sample Average Approximation and the nonparametric machine learning technique to solve the newsvendor problem with target service level constraint that is faced by the front distribution center of e-commerce enterprises. Then numerical experiments based on the real dataset of a large e-commerce enterprise are conducted to compare the performances implemented by our approach and those implemented by other well-established methods. We found that our approach can get lower surplus inventory levels while realizing higher service levels especially when the target service level is higher than 80%, which provides practical guidance for the inventory decision of the e-commerce enterprise's front distribution center under the big data environment.

Keywords Inventory · Newsvendor problem · Service level · New data-driven method · Nonparametric machine learning

1 Introduction

The advancement of the Internet promotes the rapid development of e-commerce. Consumers nowadays shop on the e-commerce platforms, leaving many different types of data such as member information, browsing history and merchant comments. Meanwhile, platforms can also record the sales, historical prices and discounts of products. With the abundant data, e-commerce enterprises can optimize the back-end

Y. Ye (✉) · X. Yang
School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China
e-mail: 18120599@bjtu.edu.cn

X. Yang
e-mail: 17113145@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_69

services such as warehousing and distribution, so as to provide consumers with high-level logistics services to cope with the fierce competition in the e-commerce market. For example, the "anticipatory shipping" adopted by many e-commerce platforms is based on the prediction of customer demand with the data of customer browsing history, shopping cart lists and so on. The predicted goods are sent to the nearest front distribution center (FDC) to the customers before they place the orders. Once orders are placed, the goods are immediately delivered to the customers, which greatly reduces the transportation time and improves customer satisfaction.

Although the concept of FDC has not been clearly defined in the academic field and the interpretations of FDC are different among various e-commerce companies, the commonality of all the FDCs is setting the warehouse nearby consumers to shorten distribution time and further to improve customer satisfaction. The central distribution center (CDC) is responsible for the FDC's inventory replenishment and transshipment. If a product purchased by a customer are in stock in FDC, then it can be delivered directly to the customer from FDC. If the product are out of stock in FDC, it has to be delivered from CDC, which not only increases the transportation time and order cost, but also results in sales loss. Therefore, it is expected that the order fulfillment rate of the FDC is as high as possible. What's more, decision makers in e-commerce enterprises usually give their specific requirements on the FDC's inventory performance indicators, such as stock-out probability, which is also defined as inventory service level. At the same time, due to the limitation of capital, the capacity of a FDC is usually much smaller than that of a CDC. Hence it is important for a FDC to determine which item and how many of it should be placed to avoid warehouse explosion and satisfy the target inventory service level performance as the same time.

The above problem is essentially a combination of demand forecasting and inventory optimization. In practice, many factors can lead to changes in consumer behavior and demand patterns. For example, promotion activities on e-commerce platforms are becoming increasingly frequent, which makes the demand distribution change dynamically. Therefore, the traditional model-driven method (i.e., assuming that the demand is subject to a specific distribution, and then solving the inventory optimization problem) is not well suitable for the inventory management in the big data era. In line with the trend of big data, research in the field of data-driven inventory management also show new opportunities and challenges. A data-driven method does not need to make assumptions on demand distribution, and it directly extracts historical observations from the demand of unknown distribution to make inventory decision. Different from the traditional data-driven approaches, new data-driven methods take not only historical demand observations but also feature data (i.e., demand-related data, e.g., historical prices, clickstream data) into account, which perform better than traditional data-driven approaches in certain situations [1].

This paper proposes an improved new data-driven method based on the nonparametric machine learning technique called K-nearest-neighbors regression (KNN) to solve the newsvendor problem with a service level constraint. Furthermore, numerical experiments are conducted on the real dataset of a large e-commerce enterprise to compare the performances of the proposed approach and other well-established methods. It is shown that the proposed approach outperforms in both realized service levels and surplus inventory levels.

The remainder of the paper is organized as follows. In Sect. 2, we review the literature on the inventory management problems using data-driven approaches. In Sect. 3, we describe the problem and reinterpret the existing data-driven benchmark approaches while presenting our new approach. Section 4 shows the numerical experiment and discussions of the results. Conclusions and research prospect are given in Sect. 5.

2 Literature Review

As we all know, the newsvendor model is one of the most important models in stochastic inventory theory, which has been studied extensively. Xie et al. [2]. It is assumed that the distribution and its parameters of demand are known and given in most inventory management textbooks [3]. In this section, we mainly review the literature on the inventory management problem where the distribution of demand is unknown and the inventory decision is directly derived from sample data, which is also known as data-driven inventory management.

Although most forecasting and inventory models hold the view that demand is sampled from a distribution, the more realistic situation is that the kind of demand distribution may vary. To be free from the assumption of demand distribution, a robust, distribution-free inventory model is presented to maximize the worst-case performance (for a review see [4]). Bertsimas and Thiele [5] introduce a data-driven approach that the required inventory level decision is obtained by a linear programming (LP) using historical demand instead of making assumptions on the demand distribution. Sample Average Approximation (SAA) is one of the most basic data-driven methods with a wider range of applications [6, 7]. For a newsvendor problem, given a set of historical demand observations, the optimal SAA order quantity is solved by minimizing the empirical risk (i.e. the average cost on historical demand observations). Other data-driven inventory optimization methods include stochastic gradient algorithm [8], adaptive value estimation method [9] and so on. In the past, due to the limitations of data availability and computing power [10], methods mentioned above only considered the observed historical demand samples.

Nowadays, as the cost for obtaining, storing and processing large amounts of data decreases substantially. Decision makers can take not only historical demand observations but also feature data into account for better performance [11]. In recent years, machine learning methods, due to the superiority of dealing with abundant data and high-dimensional feature data, has been drawing increasing attention within operation management. The latest research on data-driven inventory management focuses on using historical demand data and its feature data to obtain the optimal inventory decision rules by machine learning techniques.

Most of the latest references set the classic newsvendor problem with the goal of minimizing the expected cost as the natural starting point. A pioneer work that tries to apply machine learning methods into the inventory optimization problem is Ban and Rudin [1]. They develop two new data-driven algorithms which are based on empirical risk minimization and kernel optimization, to solve the classic newsvendor problem. It is proved that when the number of samples tends to infinity, the performances of the new algorithms are asymptotically optimal. Oroojlooyjadid et al. [12] propose a Deep Neural Network algorithm which works well on the real data. However, the limitation in its poor interpretability still remains. Huber et al. [13] point out that the objective function of the classical newsvendor problem based on the minimization of the expected cost is equivalent to the quantile loss function in machine learning approaches. Thus they develop two new data-driven methods based on Artificial Neural Network and Gradient Boosted Decision Tree respectively. These methods only need to input the feature data to solve the optimal order quantity directly, which is so-called Integrated Estimation and Optimization. Numerical experiments based on a large number of historical datasets of a large bread chain in Germany are conducted to evaluate the performance of new algorithms. The results show that the new data-driven algorithm outperforms traditional methods when the optimal service level is lower than 80%. Bertsimas and Kallus [14] propose five well-known nonparametric machine learning method frameworks: K-nearest-neighbors regression, kernel-based optimization, local regression, classification and regression trees and random forest. In these frameworks, machine learning methods are first applied for demand prediction and then combined with SAA to find the optimal order quantity. New methods optimize 88% of the inventory cost on a real-word dataset on average, compared with the traditional methods without considering feature data. Ban et al. [15] employ feature data of past products that are similar to new products, first conducting linear regressions to predict the demand of new products and then solve the multi-stage dynamic procurement problem by the improved scenario tree method. The numerical experiments on real data of Zara shows that the cost with new methods is 6–15% lower than that without considering feature data. Shi and Wang [16] study the inventory management problem in overseas warehouses of an online fashion retailer.

As far as we know, there are only two papers that solve the newsvendor problem with service level constraints while considering feature data. Beutel and Minner [11] assume the order quantity q is a liner function of feature data, and the original problem is transformed into a mixed integer linear programming problem. Experiments based on simulated data and actual data show that this method may result in below-target service levels. Niels et al. [18] hold the same assumption with Beutel and Minner [11], and a distributionally robust optimization method is applied to solve this problem. Experiments based on only simulated data show that this approach can effectively overachieve target service levels. However, the result of robust optimization is overly conservative, which means the inventory remains a high level.

Since research on new data-driven inventory management is in its infancy, this paper studies the newsvendor problem with target service level by combining feature data and nonparametric machine learning method, which is one of the existing research gaps.

3 Model Description

3.1 SL-SAA Model

In a classic newsvendor model, the newsvendor should decide how many newspapers to order every morning from the dealer. In the case of a service level (SL) constraint, the newsvendor should choose the order quantity q such that the random demand D is met with the target probability $1-\alpha$, where $\alpha \in (0, 1)$. The objective that decision maker focuses on is to satisfy the target service level first and then minimize the expected surplus inventory, which is given by

$$\min_{q\geq 0}\left\{\mathbb{E}(q-D)^{+} : \mathbb{P}[q \geq D] \geq 1-\alpha\right\} \tag{1}$$

where q is the order quantity, and D presents the stochastic demand. If the cumulative distribution F of D is given, then the solution to this problem is

$$q^{*}_{SL} := F^{-1}(1-\alpha) \tag{2}$$

If the demand distribution is not known, then (2) cannot be computed. Thus the SAA method can be applied as follows:

$$\min_{r,y,\gamma} \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

$$\text{s.t. } y_i \geq q - D_i \quad i = 1, \ldots, N \tag{4}$$

$$q + \gamma_i M \geq D_i \quad i = 1, \ldots, N \tag{5}$$

$$\frac{1}{N}\sum_{i=1}^{N} \gamma_i \leq \alpha \tag{6}$$

$$y_i \geq 0, \quad \gamma_i \in \{0, 1\}, \quad i = 1, \ldots, N \tag{7}$$

where y_i represents $(q-D_i)^{+}$, M is a large positive constant, for example, the largest number in the demand observations. N is the number of the observations. Indeed, (5) represents that when $\gamma_i = 0$, demand can be satisfied. Furthermore, (6) implies the probability that demand cannot be met is no more than α. This method is what we call SL-SAA.

3.2 LN-SAA Model

Beutel and Minner [11] impose that the relationship between the decision rule q and its explanatory variables $\boldsymbol{x}$ is linear, which means $q(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^p$, thus the problem can be reformulated as

$$\min_{\boldsymbol{r},y,\gamma} \frac{1}{N} \sum_{i=1}^{N} y_i \tag{8}$$

$$\text{s.t. } y_i \geq \boldsymbol{r}^\top \boldsymbol{x}_i - D_i \quad i = 1, \ldots, N \tag{9}$$

$$\boldsymbol{r}^\top \boldsymbol{x}_i + \gamma_i M \geq D_i \quad i = 1, \ldots, N \tag{10}$$

$$\frac{1}{N} \sum_{i=1}^{N} \gamma_i \leq \alpha \tag{11}$$

$$y_i \geq 0, \quad \gamma_i \in \{0, 1\}, \quad i = 1, \ldots, N \tag{12}$$

Then this problem is transformed into finding the optimal value of $\boldsymbol{r}^\top$. This method is what we call LN-SAA.

3.3 KNN-SAA Model

However, in practice, the relationship between feature data and decision rule may not be linear. On the basis of the nonparametric machine learning frameworks that are proposed by Bertsimas and Kallus [14], we develop a KNN-SAA approach to solve the newsvendor problem with service level constraint.

We assume that a decision maker has access to the historical data set $S_N = \{(d_1, \boldsymbol{x}_1), (d_2, \boldsymbol{x}_2), \ldots, (d_N, \boldsymbol{x}_N)\}$ and the input feature data set $X_N = \{\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+2}, \ldots, \boldsymbol{x}_{N+T}\}$, where d_i ($i = 1, \ldots, N$) is the demand in period i and $\boldsymbol{x}_i$ ($i = 1, \ldots, N$) is a vector of features in the past period i. Similar with $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ ($j = N + 1, \ldots, N + T$) is a vector of features in the future period j. We need to determine the order quantity that satisfies the target service level with feature data. Indeed, the problem is to optimize the conditional expected surplus inventory function with a service level constraint:

$$min\mathbb{E}[y(q, D)|\boldsymbol{x}] \tag{13}$$

$$\text{s.t. } \mathbb{P}[y(q, D) \geq D|\boldsymbol{x}] \geq 1 - \alpha \tag{14}$$

where $y(q, D)$ represent the surplus inventory $(q - D)^+$, α means the probability that demand cannot be met, and, obviously, $1 - \alpha$ is the target service level.

In the KNN-SAA model, (17) is approximated by minimizing the weighted sample average as

$$\min \hat{\mathbb{E}}[y(q, D)|\boldsymbol{x}] = \sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) y_i \tag{15}$$

to estimate the conditional expectation surplus inventory where y_i means the sample surplus inventory $(q - D_i)^+$. What's more, $\sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) = 1$, and it is assigned that $w(\boldsymbol{x}, \boldsymbol{x}_i) = 1/k$ if $i \in N_k(x)$ where $N_k(x) = \left\{\sum_{i=1}^{N} \mathbb{I}\left[\|x - x_i\| \geq \left\|x - x_j\right\|\right] \leq k,\ i = 1, \ldots, N\right\}$ is the neighborhood of the k nearest historical samples to the new input data $\boldsymbol{x}$ (for more details about KNN, see Altman [17]). Otherwise, $w(\boldsymbol{x}, \boldsymbol{x}_i) = 0$.

Therefore, the KNN-SAA model is formulated as follows:

$$\min \sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) y_i \tag{16}$$

$$\text{s.t. } \sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) \mathbf{1}[y(q, D) \geq 0] \geq 1 - \alpha \tag{17}$$

which is equivalent to

$$\min_{q, y, \gamma} \sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) y_i \tag{18}$$

$$\text{s.t. } y_i \geq q - D_i \quad i = 1, \ldots, N \tag{19}$$

$$q + \gamma_i M \geq D_i \quad i = 1, \ldots, N \tag{20}$$

$$\sum_{i=1}^{N} w(\boldsymbol{x}, \boldsymbol{x}_i) \gamma_i \leq \alpha \tag{21}$$

$$q \geq 0, \quad y_i \geq 0, \quad \gamma_i \in \{0, 1\}, \quad i = 1, \ldots, N \tag{22}$$

Actually, when K = N, (18)–(22) is equivalent to (3)–(7). The KNN-SAA model is a tractable mixed integer problem (MIP) and can be solved by mathematical programming solvers (e.g. CPLEX).

4 Numerical Experiment

4.1 Data Source

The data, after data masking (i.e. the technology that realize the protection of sensitive and privacy data), consist of 943 commodities that were sold on a large e-commerce platform in China and delivered from a distribution center. The time span of the dataset is from 2016-01-01 to 2017-10-31. As China's largest e-commerce holidays (i.e. "Double 11" and "6.18") are hold in November and June every year, the sales patterns in November and June are very different from those in other months. The data in November and June are excluded. Thus the dataset has a total time span of 580 days.

The information provided is divided into three parts. First, as is shown in Table 1, the sales information. Second, basic information of SKUs (see Table 2). Third, promotion information of SKUs in Table 3.

Using the dataset, we hope to extract and enrich the external explanatory features, and apply our KNN-SAA algorithm to simulate and evaluate the performance of the newsvendor problem with service level constraints, so as to determine reasonable order quantities for the distribution center.

Table 1 Sales information

Column name	Sample data	Description
item_sku_id	637	Item unique identification number
date	2016/10/12	Date
quantity	5	Sales quantity of the day
vendibility	1	Stock availability at the end of the day, 0 means no inventory left, otherwise 1
original_price	0. 0898862	Original price
discount	7.858136	Daily average discount = daily average transaction price/original price. Range is from 0 to 10, 9.5 means 5% discount

Table 2 SKUs information

Column name	Sample data	Description
item_sku_id	1	Item unique identification number
item_first_cate_cd	1	Item first level category code, e.g. shoes
item_second_cate_cd	5	Item second level category code, e.g. sports shoes
item_third_cate_cd	366	Item third level category code, e.g. running shoes
brand_code	198	Item brand code

Table 3 Promotion information

Column name	Sample data	Description
item_sku_id	1	Item unique identification number
item_third_cate_cd	366	Item 3rd level category id
date	2017/12/25	Date of promotion
promotion_type	10	A specific promotion type: e.g. direct discount, coupon, etc.

4.2 Experimental Design

1. *Data Selection and Processing*

In order to reduce the complexity of the experiment, we consider the case of single product. We randomly select one item with full selling days (i.e. 580 days), whose item_sku_id is 183. First, there is no missing value in the data of item 183. Second, as we can see from Fig. 1, the original demand of item 183, printed in blue line, is nonstationary. In order to avoid the influence of outliers, we replace the values higher than 95th percentile of sales with 95th percentile of sales, and replace the values lower than 5th percentile of sales with 5th percentile of sales. We show the processed time series sales in orange line of Fig. 1. In addition, censored demand observations only account for 2%, which indicates that 98% of the sales are true demand observations. Therefore we do not deal with those censored data.

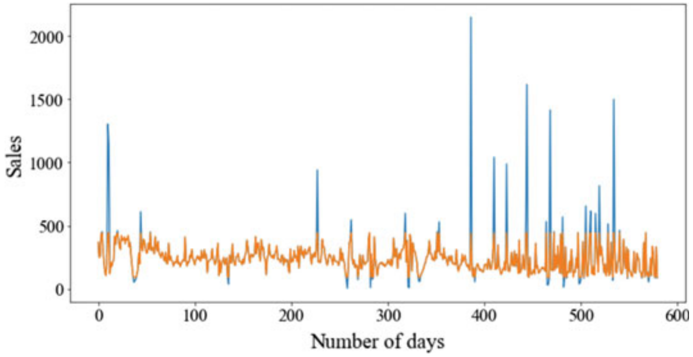


Fig. 1 Daliy sales of SKU183

Table 4 Feature generation

Data source	Features
Transactional data	Original price, discount, actual price, promotion type, lagged sales, rolling median of sales
Calendar	Year, day of year, month, day of month, day of week, public holiday

2. *Feature Generation*

Based on the dataset, we enrich the several types of features as shown in Table 4. Then the discrete feature data are conducted with one-hot encoding, and the continuous feature data are standardized. Finally, we get 40 features input.

3. *KNN-SAA Setup*

We use the data from January 2016 to July 2017 for training, those from August 2017 for validation, those from September 2017 and October 2017 for testing. Table 5 shows the lengths of training set, validation set and test set.

The validation set is used to find a reasonable value of K such that the surplus inventory level in the validation set is minimized while the realized service level is on-target or even over-target. For example, Fig. 2 shows the change of realized service level with the value of K changing in the validation set when the target service level is 0.6. It is evident that no matter what value K takes, the corresponding realized service level surpasses its target service level. However, Fig. 3 indicates that

Table 5 Sample size

Sample	Length (days)
Training set	488
Validation set	31
Test set	61

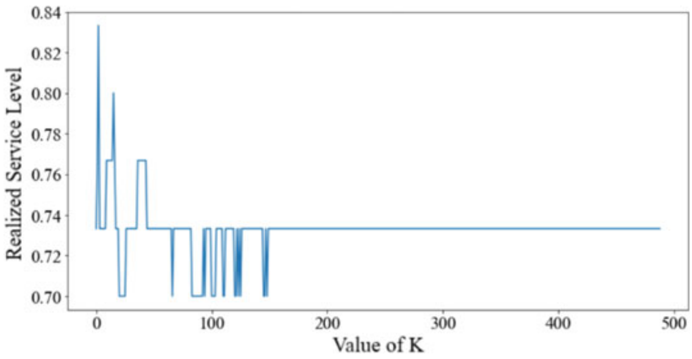


Fig. 2 Changes of realized service level with the value of K changing

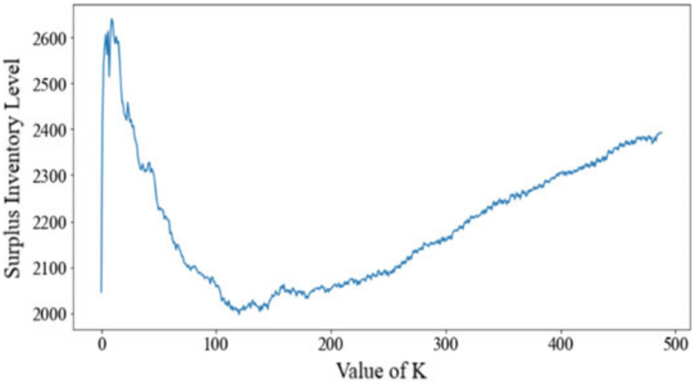


Fig. 3 Changes of surplus inventory level with the value of K changing

the minimal surplus inventory is obtained on the validation set when the value of K is 121. Hence we take 121 as the value of K on the test set.

Given that in practice, the service level is nearly always no less than 50% [12], we evaluate the performances on KNN-SAA with target service levels of 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% and 100% on test set respectively. In addition, we also implement SL-SAA and LN-SAA, to compare their performances with KNN-SAA's. All the algorithms are implemented by Python 3.7 and IBM ILOG CEPLEX 12.6.

4.3 Results

The realized service levels and surplus inventories for each approach in the cases of different target service levels are reported in Table 6. The best inventory performance for each target service level constraint is marked with an asterisk. Realized service levels that are below-target are printed in bold face. SL is the abbreviation of service level.

From Table 6, it is revealed that, in general, considering feature data helps to improve the inventory performance under certain conditions. First of all, the best inventory performance for each target service level constraint (i.e. values marked with asterisks in Table 6) is implemented by KNN-SAA or LN-SAA. LN-SAA gets the best surplus inventory performances when target service level ranges from 50 to 65%, and KNN-SAA outperforms when target service level ranges from 70 to 100%. Second, whatever the target service level is, the corresponding service level realized by KNN-SAA method is higher than that by SL-SAA method, while the corresponding surplus inventory level of KNN-SAA is lower than that of SL-SAA.

Table 6 Results

Target SL (%)	KNN-SAA		LN-SAA		SL-SAA	
	Realized SL (%)	Surplus inventory	Realized SL (%)	Surplus inventory	Realized SL (%)	Surplus inventory
50	67	3398	59	*2989	62	3727
55	67	3731	61	*3430	64	4147
60	71	4162	62	*3830	67	4590
65	74	4658	67	*4467	69	5125
70	75	*5171	74	5698	74	5563
75	80	*5665	77	7176	75	6243
80	85	*6119	**79**	7990	**79**	7462
85	92	*7516	**83**	9783	85	8782
90	93	*7921	90	11,360	**88**	10,920
95	98	*9737	**91**	12,083	**90**	13,931
100	100	*12,147	**92**	14,435	100	14,491

1. *Realized Service Level Performance Analysis*

It is presented in Fig. 4 that our KNN-SAA method can overachieve all the target service levels, which means that KNN-SAA is more reliable in achieving target service levels. Moreover, when target service level is lower than 75%, the service levels realized by LN-SAA and SL-SAA can also overachieve the target service level. But as the target service level becomes higher, the actual service levels realized by LN-SAA and SL-SAA seem more likely to be below-target. Especially when the

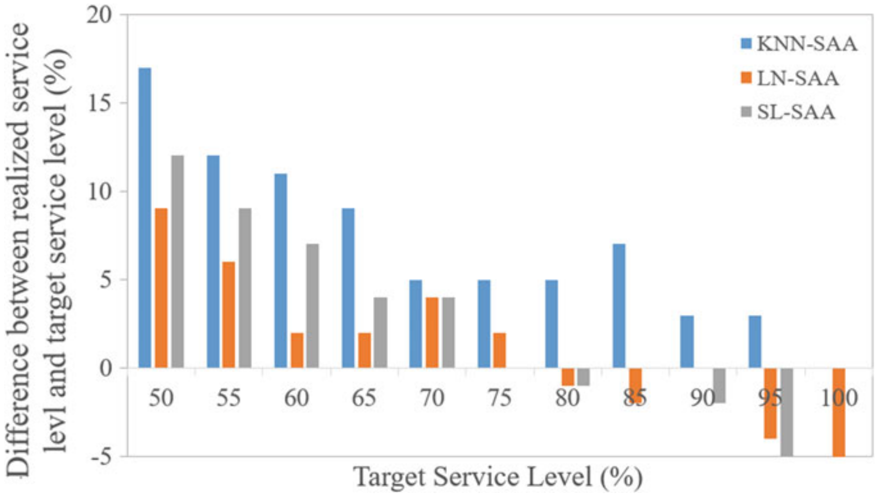


Fig. 4 Realized service level performances

target service level is 95%, the service level achieved by LN-SAA and SL-SAA are only 91% and 90%, respectively. By contrast, service level implemented by KNN-SAA is 98%. Therefore, KNN-SAA is more applicable for solving the newsvendor problem with target service level that is higher than 70%.

2. *Surplus Inventory Performance Analysis*

First of all, as is described in Fig. 5, it is evident that there is a positive correlation between target service level and its corresponding surplus inventory. Then, we take the surplus inventory performance of SL-SAA method as the baseline to compare with that of KNN-SAA and LN-SAA. Figure 6 shows that when the target service level is in [50%, 65%], the average improvement rate of surplus inventory level realized by

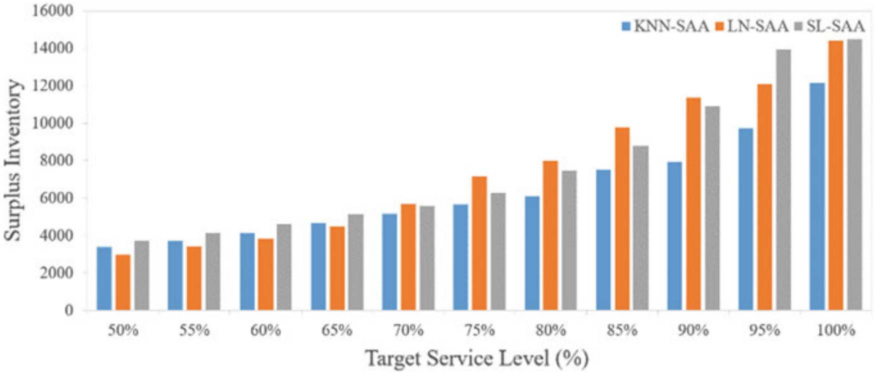


Fig. 5 Surplus inventory performances

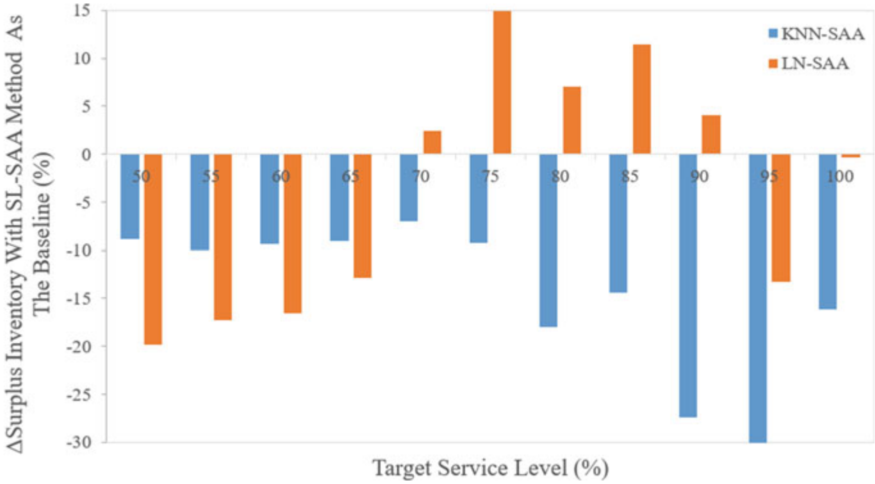


Fig. 6 Surplus inventory performances compared to the baseline method

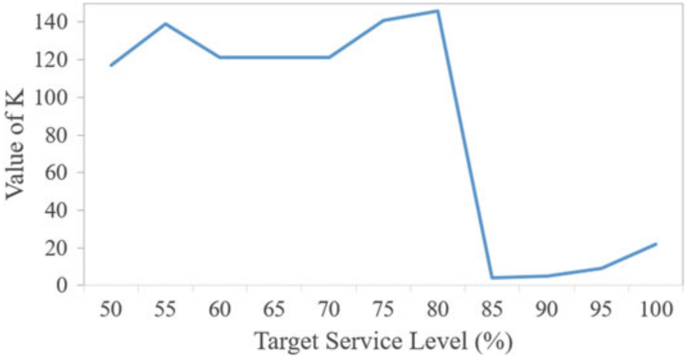


Fig. 7 Changes of the value of K with target service level changing

LN-SAA and KNN-SAA are only 16.6% and 9.3%, respectively. As the target service level increases to [70%, 100%], the average improvement rate of surplus inventory level implemented by KNN-SAA rises to 17.5%. With a service level constraint of 95%, surplus inventory level realized by KNN-SAA is even improved by 30%. Nevertheless, the average improvement rate of surplus inventory level implemented by LN-SAA falls to −3.7%, which means in this range of target service level, the average surplus inventory realized by LN-SAA performs even worse than that by SL-SAA. This phenomenon suggests that the linear assumption cannot fully interpret the relationship between order quantity q and feature data $\boldsymbol{x}$.

3. *The Effect of K*

From Fig. 7, it is deduced that the performance of KNN-SAA does not always increase in K. As is mentioned in Sect. 3, when the value of K equals to the number of historical samples, KNN-SAA model is equivalent with SL-SAA model. However, combining the results discussed above, it shows that even when K = 4 (target service level = 85%), compared with SL-SAA, KNN-SAA method can increase its realized service level by 7% while decreasing its surplus inventory by nearly 15%. Moreover, it improves the performances of both inventory and realized service level while reducing computation scales effectively.

5 Conclusion

This paper studies the inventory problem with target service level constraint faced by the front distribution center of e-commerce enterprises in the era of big data and proposes an improved new data-driven method to solve this problem. Some managerial insights are got from the results of numerical experiments on real data. First, considering feature data helps to improve the inventory performance under certain conditions. Second, KNN-SAA method is more applicable for solving the newsvendor problem with target service level higher than 70%, which, on the one hand, helps enterprises get lower surplus inventory levels while realizing higher service levels, and on the other hand, strengthens the competition force of enterprises.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under grant numbers 71962026 and 71661167009, NSFC/RGC under grant number N_PolyU531/16.

References

1. G.Y. Ban, C. Rudin, The big data newsvendor: practical insights from machine learning. Oper. Res. **67**(1), 90–108 (2019)
2. Y. Xie, L. Xiang, H.W. Wang, S. Chen Yong, An extended newsvendor model with service level constraint and capacity constraint. Ind. Eng. J. **15**(4), 47–52 (2012)
3. E.A. Silver, D.F. Pyke, D.J. Thomas, *Inventory and Production Management in Supply Chain*, 4th edn. (CRC Press, Florida, 2016)
4. G. Gallego, I. Moon, The distribution free newsboy problem: review and extensions. J. Oper. Res. Soc. **44**(8), 825–834 (1993)
5. D. Bertsimas, A. Thiele, A robust optimization approach to inventory theory. Oper. Res. **54**(1), 150–168 (2006)
6. R. Levi, G. Perakis, J. Uichanco, The data-driven newsvendor problem: new bounds and insights. Oper. Res. **63**(6), 1294–1306 (2015)
7. R. Levi, R.O. Roundy, D.B. Shmoys, Provably near-optimal sampling-based policies for stochastic inventory control models. Math. Oper. Res. **32**(4), 821–839 (2007)
8. A. Burnetas, A, C.E. Smith, Adaptive ordering and pricing for perishable products. Oper. Res. **48**(3), 436–443 (2000)
9. G.A. Godfrey, W.B. Powell, An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. Manage. Sci. **47**(8), 1101–1112 (2001)
10. V.V. Misic, G. Perakis, Data analytics in operations management: a review. Manuf. Serv. Oper. Manage. **22**(1), 158–169 (2020)
11. A.L. Beutel, S. Minner, Safety stock planning under causal demand forecasting. Int. J. Prod. Econ. **140**(2), 637–645 (2012)
12. A. Oroojlooyjadid, L.V. Snyder, M. Takác, Applying deep learning to the newsvendor problem. IISE Trans. **52**(4), 444–463 (2020)
13. J. Huber, S. Müller, M. Fleischmann, H. Stuckenschmidt, A data-driven newsvendor problem: from data to decision. Eur. J. Oper. Res. **278**(3), 904–915 (2019)
14. D. Bertsimas, N. Kallus, From predictive to prescriptive analytics. Manage. Sci. **66**(3), 1025–1044 (2020)
15. G.Y. Ban, J. Gallien, A.J. Mersereau, Dynamic procurement of new products with covariate information: the residual tree method. Manuf. Serv. Oper. Manage. **21**(4), 798–815 (2019)
16. Y. Shi, T. Wang, L.C. Alwan, Analytics for cross-border e-commerce: inventory risk management of an online fashion retailer. Decis. Sci. (2020)
17. N.S. Altman, An introduction to kernel and nearest neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
18. V.D.L Niels, R.H. Teunter, W. Romeijnders, O.A. Kilic, The data-driven newsvendor problem: achieving on-target service levels. University of Groningen, Groningen, Holland, SOM Research Reports 2019003-OPERA (2019)

Order Acceptance and Scheduling for Make-to-Order Manufacturing Enterprises

Mingzhen Yu

Abstract With limited production capacity, it is critical for make-to-order manufacturing enterprises to choose orders and make balanced scheduling to get more profits. In this paper, we study joint decision on order acceptance and scheduling problem (Order Acceptance and Scheduling, OAS). We propose OAS model considering both delay penalty and inventory penalty, which is rarely considered in past research. We develop an improved algorithm to solve this model and the experiment results verify the effectiveness of the algorithm.

Keywords Order acceptance · Scheduling · Make-to-order company · Maximize profit

1 Introduction

With the increasing customers' requirements for customization and the slowdown demand in manufacturing industry, enterprises (e.g., *Dell*, *Red Collar Group*) have adopt make-to-order production instead of make-to-stock production [1]. As shown in Fig. 1, order delivery process becomes longer and contains more uncertainty in make-to-order production. In practical manufacturing enterprises, limited production capacity and tight delivery requirements make it unrealistic to accept all the orders from customers. However, selecting orders to accept and scheduling orders to produce are executed separately by different departments. Sales departments try to get more orders without considering of production ability and detailed material supply. Discoordination of production and sales may result in tardiness or earliness penalty, even loss orders and customers' trust. For example, Land Rover got huge economic loss of tens of millions pounds because they lack of consideration for production risk [2]. To reduce internal decision-making conflicts, joint decisions on order acceptance and scheduling may keep balance between maximizing profits and maintaining stable production [3].

M. Yu (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: 14241071@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_70

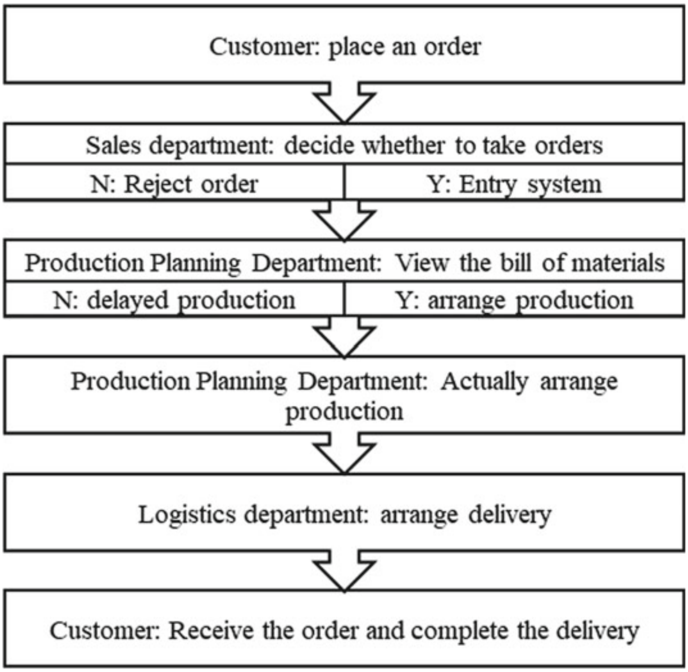

Customer: place an order
Sales department: decide whether to take orders
N: Reject order
Y: Entry system
Production Planning Department: View the bill of materials
N: delayed production
Y: arrange production
Production Planning Department: Actually arrange production
Logistics department: arrange delivery
Customer: Receive the order and complete the delivery

Fig. 1 Order delivery process for make-to-order production

In the past twenty years, OAS problem has attracted many researchers because the problem exists in real manufacturing industry, especially those make-to-order companies whose revenue is driven by orders. Palakiti [4] has concluded detailed overview in this field and basic settings contains machine workload [5], workshop capacity planning, order delivery date setting, order price and delivery date negotiation [6], etc. From the perspective of model construction, there are maximization of profits [7], minimization of the total delay time of delivery, minimization of costs, minimum total production time, and maximization the number of orders received, etc. From the perspective of solving models and algorithms, there are concluding heuristic algorithm [8], neural network algorithm, dynamic programming, queuing theory, and simulation algorithm, etc.

In conclusion, OAS problem focuses on optimizing mathematical models and designing more effective algorithms. However, there is limited researches considering practical scope. Current papers have considered lateness and punishment for delayed delivery [9], but rarely consider about inventory backlog, which also brings huge financial losses.

In this paper, we study joint decision on order acceptance and scheduling problem with the consideration of delay penalty and inventory penalty, which is realistic in many manufacturing firms. More precisely, the contribution of this paper mainly includes the following three aspects: (i) determines which orders to accept and how to scheduling the accepted orders at the same time, (ii) compares the revenue of separate decision-making and joint decision-making, (iii) provides more practical algorithms suitable for make-to-order enterprises.

The reminder of this paper is organized as follows. Section 2 presents description of the problem and model formulation. Section 3 illustrates the improved algorithm for our model. In Sect. 4, we give the computational experiments to verify the correctness of the model and the effectiveness of the algorithm and make numerical examples based on one real manufacturing company. Some conclusions and suggestions for future research is shown in Sect. 5.

2 Problem Description

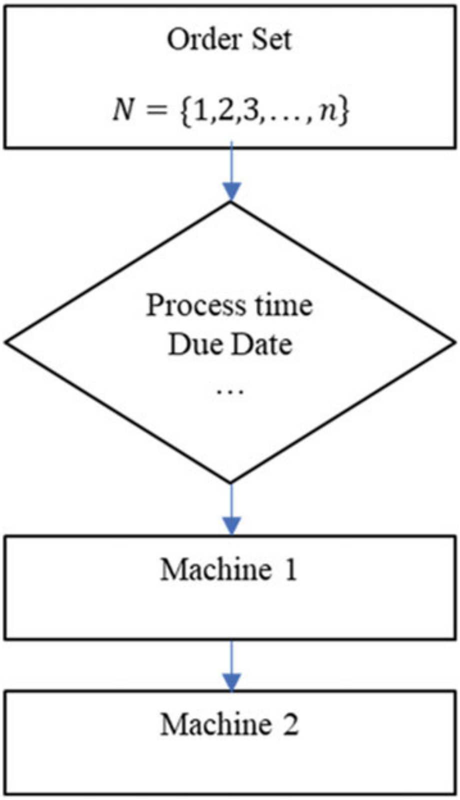
2.1 Description of OAS Problem

We describe the OAS problem that sales department and production department can share information and total order delivery process can be detected in information platform. At the beginning of order entering period, there are a set of potential customers' orders $N = \{1, 2, 3, \ldots, n\}$. We assume that company's production capacity is limited and company can accept or reject an order for processing according to the orders' delivery date, deadline and revenue. If company finishes processing an accepted order on time, they can get complete revenue without any punishments. If they produce orders earlier than given delivery date, the revenue will be equal to the complete revenue minus inventory punishment; if they produce orders later than give delivery date but earlier than deadline, the revenue will turn to be the complete revenue minus delay punishment; if they produce orders later than deadline, company will reject the orders. We assume that each machine can only process one job once a time and no job preemption is permitted. Company faces to decide which orders to accept for real processing and how to schedule the accepted jobs on machine 1 and machine 2, so as to maximize the total profits (Fig. 2).

2.2 Model Formulation

We study OAS problem on two assembly line machines, which can be expanded to multi-line machines. Total production system consists of two devices. Machine 1 is responsible for the productions of semi-finished products and Machine 2 is responsible for the productions of finished products. We assume that each order consists of the same type of reducer products, and the production processing route of each product is the same. All products in one order must be processed by the above two devices before they can be qualified for storage and waiting for shipment. Our

Fig. 2 Description of OAS problem



model aims to maximizing the total revenue of orders, considering the delay penalty and inventory costs.

Variables Definitions

Q: Real order revenue

Q_i: Real order revenue from order i

z_{ij}: Order i is accepted and processed in j position, $z_{ij} = 1$. Otherwise, $z_{ij} = 0$

w_i: Slack punishment coefficient of order i

v_i: Inventory punishment coefficient of order i

d_i: Delivery time for order i

L_i: Deadline for order i

F_i: Finish process time for order i

w_k: Waiting time for order k from process 1 to process 2

I_k: Free time between finishing order k and order $k + 1$

The specific model is as follows:

$$\text{Max } Q\sum_{j=1}^{n}\sum_{i=1}^{n} z_{ij}[Q_i - w_i(FT_i - d_i)^{+} - v_i(d_i - FT_i)^{+}] \tag{1}$$

$$\text{s.t. } \sum_{j=1}^{n} z_{jk} \le 1, \quad \forall k \in N \tag{2}$$

$$\sum_{i=1}^{n} z_{jk} \le 1, \quad \forall k \in N \tag{3}$$

$$F_i \le d_{Li}, \quad \forall i \in N \tag{4}$$

$$FT_{ij} = p_i^{(1)} + p_i^{(2)} + W_i, \quad \forall i \in N \tag{5}$$

$$W_k + \sum_{j=1}^{n} z_{jk} p_j^{(1)} + I_k - \sum_{j=1}^{n} z_{jk} p_j^{(2)} - W_{k+1} = 0, \quad \forall i \in N, j \in N, k \in N \tag{6}$$

$$z_{ij} = \begin{cases} 0, F_i > L_i \\ 1, F_i \le L_i \end{cases} \quad \forall i \in N, \quad j \in N \tag{7}$$

$$v_i = 0.001\ Q_i \quad \forall i \in N \tag{8}$$

$$v_i = \frac{1}{(d_{Li} - d_i)^{+}} \quad \forall i \in N \tag{9}$$

$$w_i = 0.001\ Q_i \quad \forall i \in N \tag{10}$$

$$w_i = \frac{1}{(d_i - d_{Li})^{+}} \quad \forall i \in N \tag{11}$$

Constraint (1) indicates that the actual benefits of orders considering delay penalty and inventory costs. Constraint (2) shows order j will be arranged at a certain processing position at most in position k. Constraint (3) means at most one order will be arranged at the k_{th} processing position. Constraint (4) means the order completion time cannot exceed the latest delivery date. Constraint (5) shows the completion time of order i is equal to the sum of the waiting time before entering machine 1, the processing time of machine 1, the waiting time before entering machine 2 after completing machine 1, and the processing time of machine 2. Constraint (6) shows there is no interference from other activity factors other than waiting for orders in production process. Constraint (7) means whether an order is accepted is a precondition for the consideration of the revenue, and rejecting the order then $z_{ij} = 0$, otherwise accept the order. Constraints (8)–(11) are inventory and slack punishment constraints.

3 Solution Procedure

According to recent research [5], the OAS problem under two-machine assembly lines is strong NP-hard. It is justified to develop effective and practical heuristics for solving OAS problem approximately. Our algorithm is based on reference [10] with considering the inventory factor. We design an improved genetic algorithm to solve our model.

The design of the improved genetic algorithm includes five parts: gene encoding and decoding, control parameters, initial population, operation steps, and fitness function.

The specific steps of the running process are shown in Fig. 3.

Step 1: Set the relevant control parameters in the genetic algorithm (population size, number of iterations, crossover rate, mutation rate, etc.).

Step 2: Set coding and decoding methods for the feasible solutions of the problem, and randomly produce initial populations.

Step 3: Establish fitness calculation function.

Step 4: Calculate the fitness value of each individual in the population.

Step 5: Judge whether the running program satisfies the stop condition, stop the operation if the condition is met, and output the corresponding solution result. Otherwise, proceed to the next step.

Step 6: Perform crossover, mutation, and selection operations to generate progeny populations.

Step 7: *Back to step 4 till getting maximum revenue.*

In the course of performing the algorithms, we take the schedule before and after performing an operation on jobs as the current schedule and the generated schedule, respectively.

4 Numerical Test

4.1 Case Background

Company A is a typical manufacturing company which uses make-to-order production to produce their famous machinery and equipment. They have more than 500,000 customers in over eighty countries, owning high quality products and reputation. They established one factory in Beijing, China, which mainly produces hydraulic pumps and gearboxes for mobile machinery. More significantly, all the operations in Beijing factory are based on their advanced Enterprise Resource Planning (ERP)

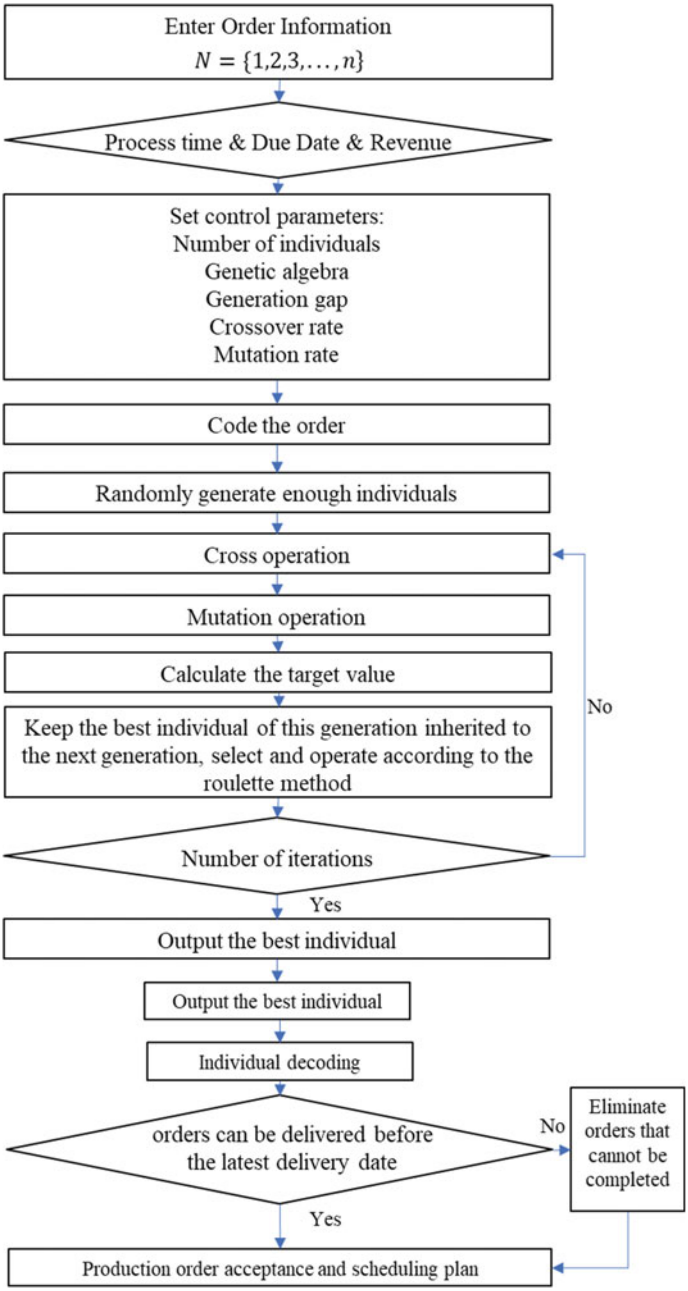

Enter Order Information
$N = \{1,2,3,\ldots,n\}$
Process time & Due Date & Revenue
Set control parameters:
Number of individuals
Genetic algebra
Generation gap
Crossover rate
Mutation rate
Code the order
Randomly generate enough individuals
Cross operation
Mutation operation
Calculate the target value
Keep the best individual of this generation inherited to the next generation, select and operate according to the roulette method
No
Number of iterations
Yes
Output the best individual
Output the best individual
Individual decoding
orders can be delivered before the latest delivery date
No
Eliminate orders that cannot be completed
Yes
Production order acceptance and scheduling plan

Fig. 3 Genetic algorithm steps

information management system, including customer order processing and production management, financial management, supply chain management and other functional modules. The system makes it possible for sales department and production department to make joint decision on order acceptance and scheduling problem. As shown in Fig. 4, they can detect their production flow and give material supply on time, making sure that there is no interruption during the production process.

As shown in Fig. 5, company A also has limited production capacity although they are the elite in the industry. Based on the realistic background of a make-to-order manufacturing company A, we strive to establish a joint decision-making model for order acceptance and production scheduling, and comprehensively consider the performance of production and sales. When the order enters, the decision-making

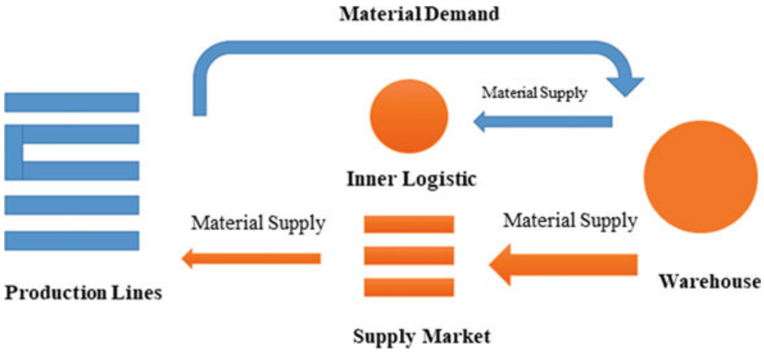


Fig. 4 A company's internal production model

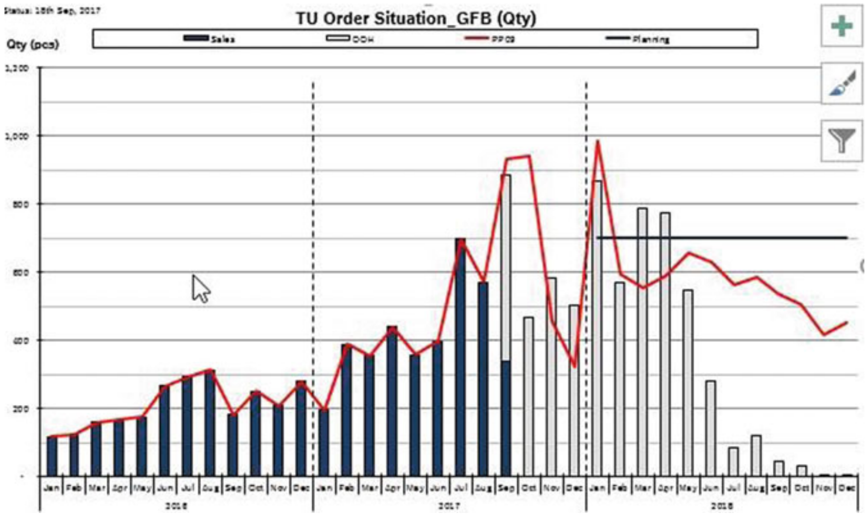


Fig. 5 Proportion of production capacity and orders

system can make arrangements and considerations for the actual production and delivery in the future, and sort the orders to integrate the production and sales coordination contradictions in actual operations. The optimization of joint decision-making is conducive to improving the enterprise's order-making decision-making strategy, scientifically carrying out coordination among internal organizations, improving the company's benchmark position in the industry, and maximizing revenue.

4.2 Numerical Test

We get data from one of company A's motor production line from January to September, 2017. In order to keep confidentiality agreement, we have desensitized the data and simulate the data according to the data regulation.

We generated 15 orders with known bill of materials, order product quantity, order processing time during machine 1 and machine 2 and original revenue promised by customers, the detail is shown in Table 1.

To test our algorithm, we conduct computational experiments to test the performance of the proposed improved genetic algorithm. We implement our algorithm with Visual studio community 2019 and IBM ILOG CPLEX Optimization Studio (64 bit) 12.8.0 on a PC computer with a double-core 2.11 GHz Intel processor and

Table 1 Orignial order revenue from sales department

Number	Material	Qty	Process time 1	Process time 2	Total revenue
1	RA	8	4.80	7.20	2894.48
2	RA	8	2.80	4.20	2894.48
3	RA	4	2.80	4.20	1447.24
4	RA	8	3.20	4.80	2894.48
5	RA	4	3.20	4.80	1447.24
6	RB	6	6.00	9.00	5430.60
7	RC	8	6.80	10.20	3105.60
8	RC	1	6.80	10.20	388.20
9	RC	1	6.00	9.00	388.20
10	RD	2	6.00	9.00	401.20
11	RE	2	6.40	9.60	360.68
12	RE	8	7.20	10.80	1442.72
13	RE	2	6.40	9.60	360.68
14	RE	8	7.20	10.80	1442.72
15	RF	1	2.40	3.60	180.34
TOTAL	/	71	78.00	117.00	25,078.86

Table 2 Orignial order revenue after production

Number	Qty	Finish time	Slack	Stock	Total revenue
1	8	4.320	0.000	0.768	2882.960
2	8	4.490	1.251	0.000	2888.230
3	4	5.740	0.000	0.126	1441.460
4	8	5.590	1.126	0.000	2888.230
5	4	6.740	0.000	0.126	1441.460
6	6	33.760	1.062	0.000	4423.980
7	8	14.490	0.000	0.251	3099.350
8	1	16.690	0.000	0.021	382.570
9	1	14.690	0.000	0.021	382.570
10	2	15.680	0.829	0.000	200.920
11	2	15.370	0.000	0.063	355.020
12	8	15.490	0.000	0.251	1436.470
13	2	15.370	0.126	0.000	355.020
14	8	15.490	0.000	0.251	1436.470
15	1	5.690	0.000	0.031	174.710
TOTAL	71	189.6	4.394	1.909	23,789.42

16.0G RAM. We provide the experimental schemes and discuss the experimental results in the following.

Formerly, sales department will accept all the fifteen orders and we run the results that many orders cannot be arranged delivery on time, causing a lot of slack and stock punishments. Once the quantity of orders increased, company A will get considerable losses. The experimental results with problem of 15 jobs are shown in Table 2.

The result shows that joint decision on OAS problem can help companies to get more revenue with considerable orders, also reducing production pressure. The experimental results with problem of 15 jobs are shown in Table 3. The experimental results show that the improved genetic algorithm is effective to solve our model and can find good solutions for our model setting efficiently.

5 Conclusion

In this study, order acceptance and scheduling play a very important role in the decision-making of manufacturing companies. If companies can apply this idea to reject some orders that bring financial loss to the company or increase the pressure of the company’s reputation, they will maximize revenue while achieving the trust of customers and reducing the pressure on suppliers. Overall, this article has made some contributions in the following areas:

Table 3 Order revenue after OAS

Number	Qty	Accept	Finish time	Slack	Stock	Total revenue
1	8	1	4.320	0.000	0.768	2888.580
2	8	1	4.490	0.000	0.251	2893.850
3	4	1	5.740	0.000	0.126	1447.080
4	8	1	5.590	0.000	0.251	2893.850
5	4	1	6.740	0.000	0.126	1447.080
6	6	1	16.880	0.531	0.000	5429.600
7	8	1	14.490	0.000	0.251	3104.970
8	1	1	16.690	0.000	0.031	388.190
9	1	1	14.690	0.000	0.031	388.190
10	2	0	/	/	/	/
11	2	1	15.370	0.000	0.063	360.640
12	8	1	15.490	0.000	0.251	1442.090
13	2	1	15.370	0.000	0.063	360.640
14	8	1	15.490	0.000	0.251	1442.090
15	1	1	5.690	0.000	0.031	180.330
Total	69	14	157.04	0.531	2.494	24,667.18

1. Realistic significance. For individual production enterprises, the analysis of their supply chain processes helps make-to-order manufacturing enterprises intuitively recognize the bottlenecks in the enterprise supply chain, and they can make better decisions of customer orders and production capacity.
2. Theoretical significance. On the basis of existing researches, this paper completes the description of the order acceptance and production scheduling problems in the dual-machine assembly line environment in the MTO production environment with consideration of delay penalty and inventory occupancy in the production process to improve the research perspective in this field. At the same time, the existing researches mainly focus on the methods of the OAS, but few focus on solving the actual problem of the enterprise. This paper attempts to make joint decision with more practical consideration and simulates the operation based on the real data of the enterprise, and conducts numerical experiments and comparisons. The analysis confirmed that the joint decision-making of order acceptance and production scheduling can effectively optimize the overall efficiency of the enterprise and enrich the research level in this field.

However, this article also has some research deficiencies. We focus on the value judgment of the production department for capacity, but lacks of the judgment of the sales department from the perspective of commercial value. For the order entry decision, corporate goals can be considered other than profit. In the next step, the relationship between the frequency of order rejection, the importance of the customer, the customer's commercial value, and the customer's reputation evaluation to ensure the company's access to profit maximization from long term prospective can be considered as further directions. Furthermore, we study order acceptance and scheduling problem based on static environment. However, mentioned in former research, the order delivery process is longer in make-to-order production system. During this period, it is possible to get urgent orders from important customers, which may affect the order acceptance decision. Therefore, companies need to reschedule the former scheduling plan once they accept those urgent orders. It is worth considering in our future study.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under grant numbers 71661167009, and 71711530714).

References

1. C. Ogŭza, F. Salmana, Z. Yalçınb, Order acceptance and scheduling decisions in make-to-order systems. Int. J. Prod. Econ. **125**, 200–211 (2010)
2. S. Slotnick, T. Morton, Order acceptance with weighted tardiness. Comput. Oper. Res. **34**, 3029–3042 (2007)
3. D. Lei, X. Guo, A parallel neighborhood search for order acceptance and scheduling in flow shop environment. Int. J. Prod. Econ. **165**, 12–18 (2015)
4. V. Palakiti, U. Mohan, V. Ganesan, Order acceptance and scheduling: overview and complexity results. Int. J. Oper. Res. **34**, 369–386 (2019)
5. X. Wang, G. Huang., X. Hu, C. Edwin, Order acceptance and scheduling on two identical parallel machines. J. Oper. Res. Soc. **66**, 1755–1767 (2015)
6. Y. Silva, A. Subramanian, A. Pessoa, Exact and heuristic algorithms for order acceptance and scheduling with sequence-dependent setup times. Comput. Oper. Res. **90**, 142–160 (2018)
7. X. Wang, Q. Zhu, T. Cheng, Subcontracting price schemes for order acceptance and scheduling. Omega **54**, 1–10 (2015)
8. M. Kebria, G. Moslehi, N. Mollaverdi, R. Mohammad, Customer's order acceptance and scheduling to maximise total profit. Int. J. Oper. Res. **34**, 301–320 (2019)
9. S. Lin, K. Ying, Order acceptance and scheduling to maximize total net revenue in permutation flowshops with weighted tardiness. Appl. Soft Comput. **30**, 462–474 (2019)
10. L. He, A. Guijt, M. Weerdt, L. Xing, N. Smith, Order acceptance and scheduling with sequence-dependent setup times: a new memetic algorithm and benchmark of the state of the art. Comput. Ind. Eng. **138**, 106102 (2019)

Bayesian Production/Inventory Competition with Unobserved Lost Sales

Shuang He, Pujie Shi, and Jian Zhang

Abstract We study the multi-period production/inventory competition problem of two substitutable products which are operated by two manufacturers respectively. The aggregate demand distribution form of products is known but one parameter is unknown before selling season. At each period, the manufacturers first update the estimates of the unknown parameters using the historical sales data, then they check the inventory levels and make the production/inventory decisions accordingly. Finally, the aggregate demand is realized and satisfied by the available products. One manufacturer's unsatisfied demand switch to the rival and is satisfied by the rival's left products if available. The unsatisfied switching demand is lost and unobserved, the unsold products incur a holding cost. The manufacturers compete for the substitute demand by making production/inventory decisions. We model the competition problem as a dynamic Bayesian game with three-dimension state-space. We reduce the dimension of state-space under certain conditions. We prove that there exists a unique Nash equilibrium in each period and show that the produce-to policy is the equilibrium production strategy. Finally, we conduct numerical experiments to examine how the model parameters impact on the equilibrium strategy and payoffs, and to generate some managerial insights.

Keywords Production competition · Bayesian game · Nash equilibrium · Produce-up-to policy

S. He (✉)
School of Economics and Management, Beijing Jiaotong University, Beijing, China
e-mail: sh.he@bjtu.edu.cn

P. Shi · J. Zhang
School of Business Administration, Inner Mongolia University of Finance and Economics, Hohhot, China
e-mail: 1121470847@qq.com

J. Zhang
e-mail: bjtuzj@bjtu.edu.cn

S. Liu et al. (eds.), *LISS 2020*,
https://doi.org/10.1007/978-981-33-4359-7_71

1 Introduction

Yili and Mengniu are two flagship brands of dairy industry in China, which accounted for about 57% of the diary market in 2018. Since 1990s, the fast development of the economy results in a diversified demand for the dairy product in China. To meet the diversified demands, both dairy-giants extend those product range by launching new products periodically. For example, Yili and Mengniu introduced their new flavor yogurts into stores, respectively, in recent year, e.g. Durian, chocolate and vanilla flavor in 2017; cheese flavor in 2018. Because these new items have no sales data, the manufacturers are uncertain about the demands as well as the demand distributions, which makes the coordination between demand and production/inventory management more challenging. In such a case, manufacturers have to learn about the demands and make production decisions simultaneous. The ambiguous demand distribution increases the risk of stockout which usually incurs a profit loss. For example, the costs of stockouts in the European grocery industry are estimated at 400 billion euros of sales and this number reach to 7–12 billion dollars in US supermarket industry [1, 10]. The first specific problem arises from stockout is that manufacturers can estimate the future demand only base on their sales observations, as lost sale is unobservable, i.e. customers who encounter an empty shelf usually leave in silence without revealing what they want to the firms. Thus, only sales data is observed, which differ from demand in case of a stock-out, such that the estimation of future demand has to be based on censored information.

The second specific problem associates with stockout is substitute competition. Studies have shown that 34% of unsatisfied consumers refuse to wait for replenishment and often take their purchasing to the rivals [1]. For example, dairy products produced by Yili and Mengniu are substitutable for each other. That is, customers who want to buy the dairy products from Yili may switch to buy the products from Mengniu when they faced a stockout. It is well known that substitution competition is an important factor that affects manufacturers' production/inventory decisions. This is because that some unsatisfied customers of one brand may look for substitutions from other brands. Under this situation, how to manage the inventory of new products with censored demand information considering substitution competition is a crucial problem faced by these two dairy-giants.

In fact, this problem is very general. The speed of product renovation is extremely high in many other important industries, examples include fast fashion computer and smart phone. And substitution competition has been verified both in theory and real practice, see Andraski and Haedicke [2] and Schary and Christopher [17].

In this paper, we study the multi-period production/inventory competition problem of two substitutable products which are operated by two manufacturers separately. The aggregate demand distribution form of products is known but one parameter is unknown before selling season. At each period, the manufacturers update the estimates of the unknown parameters using the historical sales data first, then they check the inventory levels and make the production decisions accordingly. Finally, the aggregate demand is realized and satisfied by the available products. One manufacturer's unsatisfied demand switch to the rival and is satisfied by the rival's left products if available. The unsatisfied switching demand is lost and unobserved, the unsold products incur a holding cost. The manufacturers compete for the substitute demand by making production decisions. We model the competition problem as a dynamic Bayesian game with three-dimension state-space. We reduce the dimension of state-space under certain conditions. We prove that there exists a unique Nash equilibrium in each period and show that the produce-to policy is the equilibrium production/inventory strategy. Finally, we conduct numerical experiments to examine how the model parameters impact on the equilibrium strategy and payoffs, and to generate some managerial insights.

The rest of the paper are organized as follows: we review the literature which related to our study and identify the research gap in Sect. 2. We introduce the basic game model in Sect. 3. We conduct the dimension reduction under certain conditions in Sect. 4. We characterize the Nash equilibrium production/inventory policy for the perishable case in Sect. 5. The numerical studies are conducted and managerial insights are derived in Sect. 6. We conclude the paper and propose some topics for future study in the final section.

2 Literature Review

The first stream of literature related to our work is on the multi-period inventory game. Avsar and Baykal-Gursoy [3] first considered the infinite horizon inventory game and showed that the Nash equilibrium strategy is a stationary order-to policy if the initial inventory levels of both players are low enough. Netessine et al. [15] incorporate the consumers' backordering behaviour into model and examined how it impact on the equilibrium order quantities. Nagarajan and Rajagopalan [14] considered the problem in the correlated setting in which an aggregate demand allocated to two retailers and found that under certain conditions the retailers should ignore the substitution competition. Different types of equilibrium are reviewed by Olsen and Parker [16], they proposed conditions under which the stationary infinite-horizon equilibrium equals to the Markov perfect equilibrium. Zhang et al. [18] studied a special inventory game in which the seasonal products are purchased only in the first period and then retrieved for sale in following periods. These works assume that the demand in all period are known before selling season and lost sales can be observed.

The second stream related to our research concerns the Bayesian inventory management. Lariviere and Porteus [11] showed that the optimal solution can be expressed in scalable form when demands follow the newsvendor distribution. Ding et al. [9] investigated how the demand censoring impacts on the optimal policy in newsvendor setting. Bisi and Dada [5] considered a finite-horizon newsvendor model with unobservable lost sales and proposed a determining optimal ordering and pricing policies. Bisi et al. [6] studied similar problem and shown that the Weibull is the only distribution under which the optimal policy can be described in the scalable form. Mersereau [13] investigated the impact of the inventory record inaccuracy on demand estimation considering unobserved lost sales and proposed a heuristic prescription. Chen and Chao [8] considered a multiple products inventory problem with stockout substitution in which the firm does not knows both the demand distribution of each product and the substitution probabilities between products. The above works on Bayesian inventory consider a monopoly firm with unobserved lost sales. In contracts, we consider stockout-based competition between two manufacturers.

3 Basic Model

In this section, we consider a T-period production/inventory game in which two substitutable products are produced and sold by two manufacturers, respectively. At the beginning of each period, the manufacturers check their inventory levels first and then make their production/inventory decisions accordingly. Define $I_{1,t}$ and $I_{2,t}$ as the initial inventory levels of manufacturers 1 and 2, respectively, before production in period $t\,(t = 1, 2, \ldots, t)$. Let $(q_{1,t}, q_{2,t})$ be the production/inventory quantities taken by the manufacturers. We assume that the setup lead time is zero and the production are completed immediately. Let $(y_{1,t}, y_{2,t})$ be the inventory levels after production, then we have

$$y_{1,t}{=}I_{1,t} + q_{1,t}, \, y_{2,t} = I_{2,t} + q_{2,t}.$$

Let r_i and w_i denote the selling price and unit production cost of product $i\,(i = 1, 2)$, respectively. To avoid triviality, we assume $0 < c_i < p_i$, $i = 1, 2$. At the end of each period (except period T), a unit leftover product i incurs a holding cost h_i, and each unsatisfied first-choice demand incurs a penalty cost p_i. Note that the unsatisfied substitute demand does not incur penalty cost, $i = 1, 2$.

Denote X_t as the total industry demand which is splitted to two manufacturers in period t. We introduce the Rule 1 in Lippman and McCardle [12] to allocate the industry demand in which manufacturer 1's share is γX_t and that of manufacturer 2 is $(1-\gamma)X_t, \gamma \in (0, 1)$. When one manufacturer's product is sold out, the unsatisfied customers switch to buy the product of the other manufacturer, if available, otherwise, they are lost. For example, if manufacturer 1 cannot satisfy the first-choice demand γX_t fully, all the unsatisfied demand $(\gamma X_t - y_{1,t})$ will switch to manufacturer 2 and be satisfied if it has leftover stocks, and vice versa. Thus, the effective demands $(D_{1,t}, D_{2,t})$ faced by the manufacturers are include the first-choice demand and substitute demand, and can be expressed as follows

$$\begin{aligned} D_{1,t} &= \gamma X_t + ((1-\gamma)X_t - y_{2,t})^+, \\ D_{2,t} &= (1-\gamma)X_t{+}(\gamma X_t - y_{1,t})^+. \end{aligned}$$

where $(x)^+ = \max\{x, 0\}$. Then the state transition equations are given as $I_{1,t+1} = (y_{1,t} - D_{1,t})^+$ and $I_{2,t+1} = (y_{2,t} - D_{2,t})^+$.

We assume that $\{X_1, X_2, \ldots, X_t\}$ are independently and identically distributed (i.i.d.) random variables with nonnegative continue distribution $G(X|\theta)$ in which $[\theta]$ is the fixed unknown parameter. Let $f(\theta)$ be the known prior density which represents their belief about the underlying demand distribution. For tractability, we restrict our attention to the distribution families with a sufficient statistic of fixed dimensions for a sample of independent observations. The updated probability density function of industry demand is given by:

$$g(X) = \int_{-\infty}^{\infty} g(X|\theta)f(\theta)d(\theta)$$

In each period, only sales (censored demand) can be observed by which the manufacturers form the posterior densities on the unknown parameters and it becomes the effective prior at the beginning of the next period. We assume that the distribution of industry demand is comes from the newsvendor distribution family which identified by Braden and Freimer [7]. They present several distributions, i.e., the Weibull, that included in these distribution family. This is a common assumption in many Bayesian inventory literatures, i.e., Porteus (1999), Mersereau [13], Chen (2020). Under such assumption, the distribution of X has the following form:

$$G(X|\theta) = 1 - e^{-\theta d(X)}$$

where $d(x)$ is a positive, differentiable, and increasing function for any $x \geq 0$.

For all newsvendor distributions, the gamma is a conjugate prior of them. Let a_t and S_t denote the shape and scale parameters of the gamma distribution in period t, respectively. Then the underlying density is given by

$$g(X|\theta) = \theta d'(X)e^{-\theta d(X)},$$

the prior density on the unknown parameter θ is given by

$$f(\theta|a_t, S_t) = \frac{S_t^{a_t}\theta^{a_t-1}e^{-S_t\theta}}{\Gamma(a_t)},$$

and the updated density is given by

$$g(X|a_t, S_t) = \frac{a_t S_t^{a_t} d'(X)}{[S_t + d(X)]^{a_t+1}}.$$

Let s_t be the aggregate sales of manufacturers and x_t be the realization of X in period t. Therefore, if manufacturers observe $s_t < y_{1,t}+y_{2,t}$ then they know $s_t = x_t$, if manufacturers observe $s_t = y_{1,t}+y_{2,t}$ then they can infer $s_t \geq x_t$. The manufacturers begin with a_1 and S_1 which denote the parameters of the prior distribution before selling season. Then parameters are updated as follows

$$S_t = S_1 + \sum_{j=2}^{t-1} d(s_j), a_t = a_1 + n_t,$$

where n_t is the number of exact observation of demands at the beginning of period t.

The manufacturers compete for the substitute demands by making production decisions. We model the problem as a Bayesian game. The expected payoffs of manufacturers from period t to the end of the horizon are

$$\pi_1(\vec{q}_t|a_t, S_t) = L_1(\vec{q}_t|a_t, S_t) + \beta C_1(I_{1,t+1}|a_{t+1}, S_{t+1}) \tag{3.1}$$

$$\pi_2(\vec{q}_t|a_t, S_t) = L_2(\vec{q}_t|a_t, S_t) + \beta C_2(I_{2,t+1}|a_{t+1}, S_{t+1}) \tag{3.2}$$

where $\vec{q}_t = (q_{1,t}, q_{2,t})$ is the decision vector in period t and

$$\begin{aligned}
L_1(\vec{q}_t|a_t, S_t) &= r_1 * \min(y_{1,t}, D_{1,t}) - p_1 * (\gamma X - y_{1,t})^+ \\
&\quad - z_1 * (y_{1,t} - D_{1,t})^+ - w_1 q_{1,t}, \\
L_2(\vec{q}_t|a_t, S_t) &= r_2 * \min(y_{2,t}, D_{2,t}) - p_2 * ((1-\gamma)X - y_{2,t})^+ \\
&\quad - z_2 * (y_{2,t} - D_{2,t})^+ - w_2 q_{2,t},
\end{aligned}$$

are manufacturer 1 and 2's current period payoffs. At the end of period T, each unsold product i is disposed at a salvage value $v_i (v_i < w_i)$. Then $z_i = h_i$ if $t \neq T$ and $z_i = -v_i$ otherwise.

$C_i(I_{i,t+1}|a_{t+1}, S_{t+1}) = \max_{y_{i,t} \geq I_{i,t}} \{\pi_1(\vec{q}_t|a_t, S_t)\}$ is manufacturers i's expected optimal profit from period $t+1$ to T under the equilibrium strategy, $i = 1, 2$. Denote $\beta (0 < \beta \leq 1)$ as the discount factor. We assume all cost and revenue information are symmetric and it is a common assumption in most of inventory competition literatures [14, 16, 18].

Given inventory levels after production $(y_{1,t}, y_{2,t})$, we discuss the demand information update process and the stochastic part of payoffs of manufacturers in the following two cases. Let π_i^m be the expected profit of manufacturer $i (i = 1, 2)$ in Case $m (m = 1, 2)$. Let $dG(X|a_t, S_t) = g(X|a_t, S_t) d(X)$.

Case 1. $\frac{y_{2,t}}{1-\gamma} < \frac{y_{1,t}}{\gamma}$. We can get $\frac{y_{2,t}}{1-\gamma} < y_{1,t} + y_{2,t} < \frac{y_{1,t}}{\gamma}$ through some simple calculation, which partitions the distribution range of X into four regions.

(i) $0 < X < \frac{y_{2,t}}{1-\gamma}$. In this subcase, we have $y_{1,t} \geq \gamma X$, $y_{1,t} \geq (1-\gamma)X$, and $y_{1,t} + y_{2,t} \geq X$, which means that manufacturers satisfy their first-choice demands fully and there is no switching customer. Thus, manufacturers observe all the aggregate demand x_t and update the parameters in following period. Then we have

$$\pi_1^1(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} + \int\limits_0^{y_{2,t}/(1-\gamma)} [-(r_1 - z_1)(y_{1,t} - \gamma X)$$

$$+ \beta C_{1,t+1}(y_{1,t} - \gamma X|a_t + 1, S_t + d(x_t))]dG(X|a_t, S_t),$$

$$\pi_2^1(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} + \int\limits_0^{y_{2,t}/(1-\gamma)} [-(r_2 - z_2)(y_{2,t} - (1-\gamma)X)$$

$$+ \beta C_{2,t+1}(y_{2,t} - (1-\gamma)X|a_t + 1, S_t + d(x_t))]dG(X|a_t, S_t).$$

(ii) $\frac{y_{2,t}}{1-\gamma} < X < y_{1,t} + y_{2,t}$. In this subcase, we get $y_{1,t} \geq \gamma X$, $y_{2,t} < (1-\gamma)X$, and $y_{1,t} + y_{2,t} \geq X$. Then only part of manufacturer 2's first-choice demand can be satisfied and the rest, $(1-\gamma)X - y_{2,t}$, becomes the substitute demand of manufacturer 1. After satisfying the direct and substitute demands, manufacturer 1 left $(y_{1,t} + y_{2,t} - X)$ of product. Note that the wholesaler 2 can observe all aggregate demand although he is stockout. This is because that the wholesaler 1 is not stockout $(y_{1,t} \geq \gamma X)$ and all information are symmetric, then the wholesaler 2 can infer the demand information from $x_t = y_{1,t}/\gamma$. Thus, we have

$$\pi_1^1(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} + \int\limits_{y_{2,t}/(1-\gamma)}^{y_{1,t}+y_{2,t}} [-(r_1 - z_1)(y_{1,t} + y_{2,t} - X)$$

$$+ \beta C_{1,t+1}(y_{1,t} + y_{2,t} - X|a_t + 1, S_t + d(x_t))]dG(X|a_t, S_t),$$

$$\pi_2^1(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} + \int\limits_{y_{2,t}/(1-\gamma)}^{y_{1,t}+y_{2,t}} [-p_2((1-\gamma)X - y_{2,t})$$

$$+ \beta C_{2,t+1}(0|a_t + 1, S_t + d(x_t))]dG(X|a_t, S_t).$$

(iii) $y_{1,t} + y_{2,t} < X < \frac{y_{1,t}}{\gamma}$. In this subcase, we get $y_{1,t} \geq \gamma X$, $y_{2,t} < (1-\gamma)X$, and $y_{1,t} + y_{2,t} < X$. The manufacturer 2 can satisfy $y_{2,t}$ of his first-choice demand and the rest, i.e., $(1-\gamma)X_t - y_{2,t}$, will switch to buy manufacturer 1's product. Manufacturer 1 sells all his products in this period and the switching customers cannot be satisfied fully. Thus, we have

$$\pi_1^1(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} + \int_{y_{1,t}+y_{2,t}}^{y_{1,t}/\gamma} \beta C_{1,t+1}(0|a_t + 1, S_t + d(x_t))dG(X|a_t, S_t),$$

$$\pi_2^1(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} + \int_{y_{1,t}+y_{2,t}}^{y_{1,t}/\gamma} [-p_2((1-\gamma)X - y_{2,t}) + \beta C_{2,t+1}(0|a_t + 1, S_t + d(x_t))]dG(X|a_t, S_t).$$

(iv) $\frac{y_{1,t}}{\gamma} < X_t$. In this subcase, we have $\gamma X > y_{1,t}$, $(1-\gamma)X > y_{2,t}$, and $y_{1,t}+y_{2,t} < X$. Both manufacturers are stockout and they only observe the sales of the two products. Let $C_i^B = C_i(0\big|a_t, S_t + d(y_{1,t} + y_{2,t}))$, $i = 1, 2$, then we have

$$\pi_1^1(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} + \int_{y_{1,t}/\gamma}^{\infty} [-p_1(\gamma X - y_{1,t}) + \beta C_{1,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t}))]dG(X|a_t, S_t),$$

$$\pi_2^1(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} + \int_{y_{1,t}/\gamma}^{\infty} [-p_2((1-\gamma)X - y_{2,t}) + \beta C_{2,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t}))]dG(X|a_t, S_t).$$

To sum up the above analysis, the expected functions of the stochastic part of payoffs in Case 1 are

$$\pi_1^1(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} - (r_1 - z_1) \int_0^{y_{2,t}/(1-\gamma)} (y_{1,t} - \gamma X)dG(X|a_t, S_t)$$
$$- (r_1 - z_1) \int_{y_{2,t}/(1-\gamma)}^{y_{1,t}+y_{2,t}} (y_{1,t} + y_{2,t} - X)dG(X|a_t, S_t)$$
$$- p_1 \int_{y_{1,t}/\gamma}^{\infty} (\gamma X - y_{1,t})dG(X|a_t, S_t)$$
$$+ \beta \int_0^{y_{2,t}/(1-\gamma)} C_{1,t+1}(y_{1,t} - \gamma X|a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)$$

$$+ \beta \int_{y_{2,t}/(1-\gamma)}^{y_{1,t}+y_{2,t}} C_{1,t+1}(y_{1,t} + y_{2,t} - X|a_t + 1, S_t + d(x_t))\,]dG(X|a_t, S_t)$$

$$+ \beta \int_{y_{1,t}+y_{2,t}}^{y_{1,t}/\gamma} C_{1,t+1}(0|a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)$$

$$+ \beta \int_{y_{1,t}/\gamma}^{\infty} C_{1,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t}))dG(X|a_t, S_t),$$

$$\pi_2^1(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} - (r_2 - z_2) \int_0^{y_{2,t}/(1-\gamma)} [y_{2,t} - (1-\gamma)X]dG(X|a_t, S_t)$$

$$- p_2 \int_{y_{2,t}/(1-\gamma)}^{\infty} [(1-\gamma)X - y_{2,t}]dG(X|a_t, S_t)$$

$$+ \beta \int_0^{y_{2,t}/(1-\gamma)} C_{2,t+1}(y_{2,t} - (1-\gamma)X|a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)$$

$$+ \beta \int_{y_{2,t}/(1-\gamma)}^{y_{1,t}/\gamma} C_{2,t+1}(0|a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)$$

$$+ \beta \int_{y_{1,t}/\gamma}^{\infty} C_{2,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t}))dG(X|a_t, S_t).$$

The first two items of π_2^1 are sales revenue and total production cost respectively. The third item of π_2^1 is the expected loss under the scenario where manufacturer 2 has surplus products after satisfying his first-choice demand (from $\frac{y_{2,t}}{1-\gamma} < \frac{y_{1,t}}{\gamma}$, manufacturer 1 is also overstocked. Then there is no interaction between the manufacturers). The fourth item of π_2^1 is the penalty cost under the scenario where manufacturer 2 cannot satisfy the first-choice demand fully.

Case 2. $\frac{y_{1,t}}{\gamma} \le \frac{y_{2,t}}{1-\gamma}$. By a similar argument as in Case 1, we can formulate the two manufacturers' expected payoff functions as follows:

$$\pi_1^2(\vec{q}_t|a_t, S_t) = r_1 y_{1,t} - w_1 q_{1,t} - (r_1 - z_1) \int_0^{y_{1,t}/\gamma} (y_{1,t} - \gamma X)dG(X|a_t, S_t)$$

$$
\begin{aligned}
&- p_1 \int\limits_{y_{1,t}/\gamma}^{\infty} (\gamma X - y_{1,t}) dG(X|a_t, S_t) \\
&+ \beta \int\limits_{0}^{y_{1,t}/\gamma} C_{1,t+1}(y_{1,t} - \gamma X|a_t + 1, S_t + d(x_t)) dG(X|a_t, S_t) \\
&+ \beta \int\limits_{y_{1,t}/\gamma}^{y_{2,t}/(1-\gamma)} C_{1,t+1}(0|a_t + 1, S_t + d(x_t)) dG(X|a_t, S_t) \\
&+ \beta \int\limits_{y_{2,t}/(1-\gamma)}^{\infty} C_{1,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t})) dG(X|a_t, S_t),
\end{aligned}
$$

$$
\begin{aligned}
\pi_2^2(\vec{q}_t|a_t, S_t) = r_2 y_{2,t} - w_2 q_{2,t} - (r_2 - z_2) \int\limits_{0}^{y_{1,t}/\gamma} [y_{2,t} - (1-\gamma)X] dG(X|a_t, S_t) \\
- (r_2 - z_2) \int\limits_{y_{1,t}/\gamma}^{y_{1,t}+y_{2,t}} [y_{1,t} + y_{2,t} - X] dG(X|a_t, S_t) \\
+ p_2 \int\limits_{y_{2,t}/(1-\gamma)}^{\infty} [(1-\gamma)X - y_{2,t}] dG(X|a_t, S_t) \\
- p_1 \int\limits_{y_{1,t}/\gamma}^{\infty} (\gamma X - y_{1,t}) dG(X|a_t, S_t) \\
+ \beta \int\limits_{0}^{y_{1,t}/\gamma} C_{2,t+1}(y_{2,t} - (1-\gamma)|a_t + 1, S_t + d(x_t)) dG(X|a_t, S_t) \\
+ \beta \int\limits_{y_{1,t}/\gamma}^{y_{1,t}+y_{2,t}} C_{2,t+1}(y_{1,t} + y_{2,t} - X|a_t + 1, S_t + d(x_t)) dG(X|a_t, S_t) \\
+ \beta \int\limits_{y_{1,t}+y_{2,t}}^{y_{2,t}/(1-\gamma)} C_{2,t+1}(0|a_t + 1, S_t + d(x_t)) dG(X|a_t, S_t) \\
+ \beta \int\limits_{y_{2,t}/(1-\gamma)}^{\infty} C_{2,t+1}(0\big|a_t, S_t + d(y_{1,t} + y_{2,t})) dG(X|a_t, S_t)
\end{aligned}
$$

Lariviere and Porteus [11] and Avsar and Baykal-Gursoys' [3] models are included in Model (3.1) and (3.2). Avsar and Baykal-Gursoy [3] studied multi-period inventory game with known demand distributions and the lost sales can be observed by firms in each period. Here we assume that the distributions of demands are unknown and unobserved lost sales in the all periods. The manufacturers have to update the estimation of demand distribution based on the historic sales data before making decisions in each period. Lariviere and Porteus [11] studied the Bayesian inventory problem in the monopolistic case. Here we generalized the model in Lariviere and Porteus [11] to the duopoly competition setting.

4 Dimensionality Reduction

The historical sales data impact manufacturers' production strategies through the parameters (a_t, S_t) then the state space of the Bayesian game includes three dimensions (initial inventory level and sales history). In this section, we adopt the dimensionality reduction technology which is presented by Scarf (1959) [19] to eliminate the parameter S_t from the analysis and give some conditions under which the Bayesian game (3.1) and (3.2) is equivalent to a low-dimension game. We define the standard case as $S_t = 1$ and suppress S_t in our model, e.g., $g(X|a_t, 1) = g(X|a_t)$, $L_i(\vec{q}_t|a_t, 1) = L_i(\vec{q}_t|a_t)$, $\pi_i(\vec{q}_t|a_t, 1) = \pi_i(\vec{q}_t|a_t)$, $i, m = 1, 2; t = 1, 2, 3 \dots T$. From Azoury (1985) [4] we have

Lemma 1 (Azoury 1985) *If the underlying demand distributions are Weibull (so that $d(X) = X^k$) with known k and unknown $\theta$and the prior on $\theta$are gamma with parameters (a_t, S_t), then, letting $q(S_t) = S_t^{1/k}$and $U(X) = (1 + X^k)^{1/k}$, the following hold.*

(a) $g(X|a_t, S_t) = g(X/q(S_t)|a_t)/q(S_t)$
(b) $q(S_t + x^k) = q(S_t)U(x/q(S_t))$
(c) $\int_0^\infty U(X)g(X|a_t)dX < \infty$

Let $\widehat{X} = X/q(S_t)$, $\widehat{I}_{i,t} = I_{i,t}/q(S_t)$, $\hat{y}_{i,t} = y_{i,t}/q(S_t)$, $\hat{q}_{i,t} = q_{i,t}/q(S_t)$. Define $\hat{\pi}_i^m(\vec{q}_t'|a_t)$, $(i, m = 1, 2; t = 1, 2, 3, \dots, T)$ as

$$\begin{aligned}\hat{\pi}_1^1(\vec{q}_t'|a_t) = r_1\hat{y}_{1,t} - w_1\hat{q}_{1,t} - (r_1 - z_1) \int_0^{\hat{y}_{2,t}/(1-\gamma)} (\hat{y}_{1,t} - \gamma\widehat{X})dG(\widehat{X}|a_t)\\ - (r_1 - z_1) \int_{\hat{y}_{2,t}/(1-\gamma)}^{\hat{y}_{1,t}+\hat{y}_{2,t}} (\hat{y}_{1,t} + \hat{y}_{2,t} - \hat{X})dG(\widehat{X}|a_t)\\ - p_1 \int_{\hat{y}_{1,t}/\gamma}^{\infty} (\gamma\widehat{X} - \hat{y}_{1,t})dG(\widehat{X}|a_t)\end{aligned}$$

$$+ \beta \int_{0}^{\hat{y}_{2,t}/(1-\gamma)} \widehat{C}_{1,t+1}\Big(\frac{\hat{y}_{1,t} - \gamma \widehat{X}}{U(\hat{X})}|a_t{+}1\Big) U(\widehat{X}) dG(\widehat{X}|a_t)$$

$$+ \beta \int_{\hat{y}_{2,t}/(1-\gamma)}^{\hat{y}_{1,t}+\hat{y}_{2,t}} \widehat{C}_{1,t+1}\Big(\frac{\hat{y}_{1,t} + \hat{y}_{2,t} - \widehat{X}}{U(\hat{X})}|a_t + 1\Big) U(\hat{X}) dG(\hat{X}|a_t)$$

$$+ \beta \int_{\hat{y}_{1,t}+\hat{y}_{2,t}}^{\hat{y}_{1,t}/\gamma} \widehat{C}_{1,t+1}(0|a_t + 1) U(\widehat{X}) dG(\widehat{X}|a_t)$$

$$+ \beta \int_{\hat{y}_{1,t}/\gamma}^{\infty} \widehat{C}_{1,t+1}(0|a_t) U(\hat{y}_{1,t} + \hat{y}_{2,t}) dG(\widehat{X}|a_t),$$

$$\hat{\pi}_2^1(\vec{q}_t^{\,\prime}|a_t) = r_2 \hat{y}_{2,t} - w_2 \hat{q}_{2,t} - (r_2 - z_2) \int_{0}^{\hat{y}_{2,t}/(1-\gamma)} [\hat{y}_{2,t} - (1-\gamma)\hat{X}] dG(\widehat{X}|a_t)$$

$$- p_2 \int_{\hat{y}_{2,t}/(1-\gamma)}^{\infty} [(1-\gamma)\hat{X} - \hat{y}_{2,t}] dG(\widehat{X}|a_t)$$

$$+ \beta \int_{0}^{\hat{y}_{2,t}/(1-\gamma)} \widehat{C}_{2,t+1}\Big(\frac{\hat{y}_{2,t} - (1-\gamma)\widehat{X}}{U(\widehat{X})}|a_t + 1\Big) U(\widehat{X}) dG(\widehat{X}|a_t)$$

$$+ \beta \int_{\hat{y}_{2,t}/(1-\gamma)}^{\hat{y}_{1,t}/\gamma} \widehat{C}_{2,t+1}(0|a_t + 1) U(\widehat{X}) dG(\widehat{X}|a_t)$$

$$+ \beta \int_{\hat{y}_{1,t}/\gamma}^{\infty} \widehat{C}_{2,t+1}(0|a_t) U(\hat{y}_{1,t} + \hat{y}_{2,t}) dG(\widehat{X}|a_t),$$

$$\hat{\pi}_1^2(\vec{q}_t^{\,\prime}|a_t) = r_1 \hat{y}_{1,t} - w_1 \hat{q}_{1,t} - (r_1 - z_1) \int_{0}^{\hat{y}_{1,t}/\gamma} (\hat{y}_{1,t} - \gamma \widehat{X}) dG(\widehat{X}|a_t)$$

$$- p_1 \int_{\hat{y}_{1,t}/\gamma}^{\infty} (\gamma \widehat{X} - \hat{y}_{1,t}) dG(\widehat{X}|a_t)$$

$$+ \beta \int_{0}^{\hat{y}_{1,t}/\gamma} \widehat{C}_{1,t+1}\Big(\frac{\hat{y}_{1,t} - \gamma \widehat{X}}{U(\widehat{X})}|a_t + 1\Big) U(\widehat{X}) dG(\widehat{X}|a_t)$$

$$
\begin{aligned}
&+\beta \int\limits_{y_{1,t}/\gamma}^{\hat{y}_{2,t}/(1-\gamma)} \widehat{C}_{1,t+1}(0|a_t+1)U(\widehat{X})dG(\widehat{X}|a_t) \\
&+\beta \int\limits_{\hat{y}_{2,t}/(1-\gamma)}^{\infty} \widehat{C}_{1,t+1}(0|a_t)U(\hat{y}_{1,t}+\hat{y}_{2,t})dG(\widehat{X}|a_t),
\end{aligned}
$$

$$
\begin{aligned}
\hat{\pi}_2^2(\vec{q}_t'|a_t) = r_2\hat{y}_{2,t} - w_2\hat{q}_{2,t} &- (r_2-z_2)\int\limits_{0}^{\hat{y}_{1,t}/\gamma}[\hat{y}_{2,t}-(1-\gamma)\widehat{X}]dG(\widehat{X}|a_t) \\
&- (r_2-z_2)\int\limits_{\hat{y}_{1,t}/\gamma}^{\hat{y}_{1,t}+\hat{y}_{2,t}}[\hat{y}_{1,t}+\hat{y}_{2,t}-\widehat{X}]dG(\widehat{X}|a_t) \\
&+ p_2\int\limits_{\hat{y}_{2,t}/(1-\gamma)}^{\infty}[(1-\gamma)\widehat{X}-\hat{y}_{2,t}]dG(\widehat{X}|a_t) \\
&+\beta\int\limits_{0}^{\hat{y}_{1,t}/\gamma}\widehat{C}_{2,t+1}(\frac{\hat{y}_{2,t}-(1-\gamma)\widehat{X}}{U(\widehat{X})}|a_t+1)U(\widehat{X})dG(\widehat{X}|a_t) \\
&+\beta\int\limits_{\hat{y}_{1,t}/\gamma}^{\hat{y}_{1,t}+\hat{y}_{2,t}}\widehat{C}_{2,t+1}(\frac{\hat{y}_{1,t}+\hat{y}_{2,t}-\widehat{X}}{U(\widehat{X})}|a_t+1)\,U(\widehat{X})dG(\widehat{X}|a_t) \\
&+\beta\int\limits_{\hat{y}_{1,t}+\hat{y}_{2,t}}^{\hat{y}_{2,t}/(1-\gamma)}\widehat{C}_{2,t+1}(0|a_t+1)U(\widehat{X})dG(\widehat{X}|a_t) \\
&+\beta\int\limits_{\hat{y}_{2,t}/(1-\gamma)}^{\infty}\widehat{C}_{2,t+1}(0|a_t)U(\hat{y}_{1,t}+\hat{y}_{2,t})dG(\widehat{X}|a_t),
\end{aligned}
$$

where $\vec{q}_t' = (\hat{q}_{1,t}, \hat{q}_{2,t})$. Let

$$
\begin{aligned}
&\widehat{C}_i(\widehat{I}_{1,t}|a_t) = q(S_t)\min_{\hat{y}_{1,t}\geq\widehat{I}_{1,t}}\left\{\hat{\pi}_i(\vec{q}_t'|a_t)\right\} \\
&i = 1, 2; t = 1, 2, 3, \ldots, T.
\end{aligned}
$$

Theorem 1 *In period t*, we have

(a) $\pi_i^m(\vec{q}_t|a_t, S_t) = q(S_t)\hat{\pi}_i^m(\vec{q}_t'|a_t)$;

(b) $C_i(I_{i,t}|a_t, S_t) = q(S_t)\widehat{C}_i(\widehat{I}_{i,t}|a_t)$, $i, m = 1, 2$.

Proof Please see the Appendix. □

5 The Equilibrium of Production Competition

In this section, we restrict our attention on the perishable products. When two substitute perishable products are product and sell by two manufacturers separately, (i.e. no product can be carried to the subsequent periods) the inventory before production at each period is zero. In such case, $z_i = -v_i$ and the total expected payoffs of manufacturers in both cases can be rewritten as

Case 1.

$$\begin{aligned}\pi_1^1(\vec{q}_t|a_t, S_t) = (r_1 - w_1)q_{1,t} - (r_1 + v_1) &\int_0^{q_{2,t}/(1-\gamma)} (q_{1,t} - \gamma X)dG(X|a_t, S_t)\\
&- (r_1 + v_1) \int_{q_{2,t}/(1-\gamma)}^{q_{1,t}+q_{2,t}} (q_{1,t} + q_{2,t} - X)dG(X|a_t, S_t)\\
&- p_1 \int_{q_{1,t}/\gamma}^{\infty} (\gamma X - q_{1,t})dG(X|a_t, S_t)\\
&+ \beta \int_0^{q_{1,t}/\gamma} C_{1,t+1}(a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)\\
&+ \beta \int_{q_{1,t}/\gamma}^{\infty} C_{1,t+1}(a_t, S_t + d(y_{1,t} + y_{2,t}))dG(X|a_t, S_t) \qquad (5.1)\end{aligned}$$

$$\begin{aligned}\pi_2^1(\vec{q}_t|a_t, S_t) = (r_2 - w_2)q_{2,t} - (r_2 + v_2) &\int_0^{q_{2,t}/(1-\gamma)} [q_{2,t} - (1 - \gamma)X]dG(X|a_t, S_t)\\
&- p_2 \int_{q_{2,t}/(1-\gamma)}^{\infty} [(1 - \gamma)X - q_{2,t}]dG(X|a_t, S_t)\\
&+ \beta \int_0^{q_{1,t}/\gamma} C_{2,t+1}(a_t + 1, S_t + d(x_t))dG(X|a_t, S_t)\end{aligned}$$

$$+ \beta \int_{q_{1,t}/\gamma}^{\infty} C_{2,t+1}(a_t, S_t + d(q_{1,t} + q_{2,t}))dG(X|a_t, S_t) \tag{5.2}$$

Case 2.

$$\begin{aligned}
\pi_1^2(\vec{q}_t|a_t, S_t) = (r_1 - w_1)q_{1,t} - (r_1 + v_1) & \int_{0}^{q_{1,t}/\gamma} (q_{1,t} - \gamma X)dG(X|a_t, S_t) \\
& - p_1 \int_{q_{1,t}/\gamma}^{\infty} (\gamma X - q_{1,t})dG(X|a_t, S_t) \\
& + \beta \int_{0}^{q_{2,t}/(1-\gamma)} C_{1,t+1}(a_t + 1, S_t + d(x_t))dG(X|a_t, S_t) \\
& + \beta \int_{q_{2,t}/(1-\gamma)}^{\infty} C_{1,t+1}(a_t, S_t + d(q_{1,t} + q_{2,t}))dG(X|a_t, S_t)
\end{aligned} \tag{5.3}$$

$$\begin{aligned}
\pi_2^2(\vec{q}_t|a_t, S_t) = (r_2 - w_2)q_{2,t} - (r_2 + v_2) & \int_{0}^{q_{1,t}/\gamma} [q_{2,t} - (1 - \gamma)X]dG(X|a_t, S_t) \\
& - (r_2 + v_2) \int_{y_{1,t}/\gamma}^{q_{1,t}+q_{2,t}} [q_{1,t} + q_{2,t} - X]dG(X|a_t, S_t) \\
& + p_2 \int_{q_{2,t}/(1-\gamma)}^{\infty} [(1 - \gamma)X - q_{2,t}]dG(X|a_t, S_t) \\
& - p_1 \int_{q_{1,t}/\gamma}^{\infty} (\gamma X - q_{1,t})dG(X|a_t, S_t) \\
& + \beta \int_{0}^{q_{2,t}/(1-\gamma)} C_{2,t+1}(a_t + 1, S_t + d(x_t))dG(X|a_t, S_t) \\
& + \beta \int_{q_{2,t}/(1-\gamma)}^{\infty} C_{2,t+1}(a_t, S_t + d(q_{1,t} + q_{2,t}))dG(X|a_t, S_t)
\end{aligned} \tag{5.4}$$

As Sect. 3, the following game should be considered.
Case 1.

$$\begin{aligned}\hat{\pi}_1^1(\vec{q}_t'|a_t) = (r_1 - w_1)\hat{q}_{1,t} - (r_1 + v_1) &\int\limits_0^{\hat{q}_{2,t}/(1-\gamma)} (\hat{q}_{1,t} - \gamma\widehat{X})dG(\widehat{X}|a_t)\\ &- (r_1 + v_1)\int\limits_{\hat{q}_{2,t}/(1-\gamma)}^{\hat{q}_{1,t}+\hat{q}_{2,t}} (\hat{q}_{1,t} + \hat{q}_{2,t} - \widehat{X})dG(\widehat{X}|a_t)\\ &- p_1\int\limits_{\hat{q}_{1,t}/\gamma}^{\infty} (\gamma\widehat{X} - \hat{q}_{1,t})dG(\widehat{X}|a_t)\\ &+ \beta\int\limits_0^{\hat{q}_{1,t}/\gamma} \widehat{C}_{1,t+1}(a_t{+}1)U(\widehat{X})dG(\widehat{X}|a_t)\\ &+ \beta\int\limits_{\hat{q}_{1,t}/\gamma}^{\infty} \widehat{C}_{1,t+1}(a_t)U(\hat{q}_{1,t} + \hat{q}_{2,t})dG(\widehat{X}|a_t) \qquad (5.5)\end{aligned}$$

$$\begin{aligned}\hat{\pi}_2^1(\vec{q}_t'|a_t) = (r_2 - w_2)\hat{q}_{2,t} - (r_2 + v_2) &\int\limits_0^{\hat{q}_{2,t}/(1-\gamma)} [\hat{q}_{2,t} - (1-\gamma)\widehat{X}]dG(\widehat{X}|a_t)\\ &- p_2\int\limits_{\hat{q}_{2,t}/(1-\gamma)}^{\infty} [(1-\gamma)\widehat{X} - \hat{q}_{2,t}]dG(\widehat{X}|a_t)\\ &+ \beta\int\limits_0^{\hat{q}_{1,t}/\gamma} \widehat{C}_{2,t+1}(a_t + 1)U(\widehat{X})dG(\hat{X}|a_t)\\ &+ \beta\int\limits_{\hat{q}_{1,t}/\gamma}^{\infty} \widehat{C}_{2,t+1}(a_t)U(\hat{q}_{1,t} + \hat{q}_{2,t})dG(\widehat{X}|a_t) \qquad (5.6)\end{aligned}$$

Case 2.

$$\begin{aligned}\hat{\pi}_1^2(\vec{q}_t'|a_t) = (r_1 - w_1)\hat{q}_{1,t} - (r_1 + v_1) &\int\limits_0^{\hat{q}_{1,t}/\gamma} (\hat{q}_{1,t} - \gamma\widehat{X})dG(\widehat{X}|a_t)\\ &- p_1\int\limits_{\hat{q}_{1,t}/\gamma}^{\infty} (\gamma\widehat{X} - \hat{q}_{1,t})dG(\widehat{X}|a_t)\end{aligned}$$

$$
\begin{aligned}
&+\beta \int_{0}^{\hat{q}_{2,t}/(1-\gamma)} \widehat{C}_{1,t+1}(a_t+1)U(\widehat{X})dG(\widehat{X}|a_t)\\
&+\beta \int_{\hat{q}_{2,t}/(1-\gamma)}^{\infty} \widehat{C}_{1,t+1}(a_t)U(\hat{q}_{1,t}+\hat{q}_{2,t})dG(\widehat{X}|a_t)
\end{aligned}
\tag{5.7}
$$

$$
\begin{aligned}
\hat{\pi}_2^2(\vec{q}_t^{\,\prime}|a_t) = {} & (r_2-w_2)\hat{q}_{2,t} - (r_2+v_2)\int_{0}^{\hat{q}_{1,t}/\gamma} [\hat{q}_{2,t}-(1-\gamma)\widehat{X}]dG(\widehat{X}|a_t)\\
&-(r_2+v_2)\int_{\hat{q}_{1,t}/\gamma}^{\hat{q}_{1,t}+\hat{q}_{2,t}} [\hat{q}_{1,t}+\hat{q}_{2,t}-\widehat{X}]dG(\widehat{X}|a_t)\\
&+p_2\int_{\hat{q}_{2,t}/(1-\gamma)}^{\infty} [(1-\gamma)\widehat{X}-\hat{q}_{2,t}]dG(\widehat{X}|a_t)\\
&+\beta \int_{0}^{\hat{q}_{2,t}/(1-\gamma)} \widehat{C}_{2,t+1}(a_t+1)U(\widehat{X})dG(\widehat{X}|a_t)\\
&+\beta \int_{\hat{q}_{2,t}/(1-\gamma)}^{\infty} \widehat{C}_{2,t+1}(a_t)U(\hat{q}_{1,t}+\hat{q}_{2,t})dG(\widehat{X}|a_t)
\end{aligned}
\tag{5.8}
$$

Note that game (5.5) and (5.8) is a dynamic game with complete information, and the payoff and action depend on history through shape parameter $a_t, t = 1, 2, \ldots, T$. In the following we focus on the symmetric case in which the manufacturers have the same costs and revenue functions, and the aggregate demand is equally allocated to the manufacturers.

Theorem 2 *In the case of an exponential underlying demand distribution and perishable inventory, for the game* (5.5)–(5.8) *we have*

(a) *It has a unique Nash equilibrium* $(\hat{q}_{1,t}^*, \hat{q}_{2,t}^*)$;
(b) *The Nash equilibrium* $(\hat{q}_{1,t}^*, \hat{q}_{2,t}^*)$ *are increasing in shape parameter* a_t;
(c) *The equilibrium payoffs* $(\widehat{C}_{1,t}^*(a_t), \widehat{C}_{2,t}^*(a_t))$ *are increasing in shape parameter* $a_t, t = 1, 2, 3, \cdots, T$.

Proof Please see the Appendix. □

From Theorem 2(b) we know that the Nash equilibrium solution is increasing in a_t. This is because a small a_t means more over stock in past periods which implies demand is less than expected. The mean, variance of the updated demand distribution decreasing in a_t which reduces the risk of understock and overstock costs. Thus, the equilibrium payoffs are increasing in a_t as shown in Theorem 2(c). Theorem 3 characterizes the Nash equilibrium strategy of game (5.4)–(5.8).

Theorem 3 *The unique Nash equilibrium for the Bayesian production/inventory game* (3.1)*and* (3.2) *is the produce-up-to policy, i.e., it is optimal for manufacturers to bring inventory levels up to* $\left\{q(S_t)\hat{q}^*_{1,t}, q(S_t)\hat{q}^*_{2,t}\right\}_{t=1}^{T}$ *by production.*

Proof From Theorem 1 we know that the Nash equilibrium of the game (5.5)–(5.8) is attained if $q^*_{1,t}/q(S_t) = \hat{q}^*_{1,t}$, $q^*_{2,t}/q(S_t) = \hat{q}^*_{2,t}$ in period t. Thus, the unique Nash equilibrium of game (3.1) and (3.2) is $(q(S_t)\hat{q}^*_{1,t}, q(S_t)\hat{q}^*_{2,t})$. Then the Nash equilibrium strategy for the manufacturers is produce-up-to policy with levels $\left\{q(S_t)\hat{q}^*_{1,t}, q(S_t)\hat{q}^*_{2,t}\right\}_{t=1}^{T}$.

6 Numerical Experiments

In this section, we conduct numerical studies to investigate the effect of competition and shape parameter a_1 on equilibrium strategy and payoff. The two manufacturers are symmetric and aggregate demand follows exponential distribution, i.e. $k = 1$, then we have $q(S_t) = S_t$, $U(X) = 1 + X$, and the update density is

$$g(X|a_t, S_t) = \frac{a_t S_t^{a_t}}{(S_t + X)^{a_t+1}}.$$

The cost parameters are: $r_i = 10, w_i = 5, v_i = 1, p_i = 1, i = 1, 2$. We first investigate the impact of substitute competition by comparing the production/inventory quantities and corresponding payoffs in competitive case with those of monopolistic case. In the monopolistic case, the product is produced and sold by only one manufacturer in the market. The results are shown in Table 1. From Table 1 we know that both the equilibrium production/inventory quantity and corresponding payoff of the competitive case are larger than those of the monopolistic case. This is because the switching demands will result in a higher total demand, which will induce manufacturers to produce more and earns additional profit from satisfying substitute demand.

Table 1 Comparison between competitive and monopolistic cases

Case	Q$_1$	Q$_2$	Q$_3$	Q$_4$	Sum	Payoff
Competition	13.8	12.2	10.1	9.6	45.7	447.86
Monopoly	11.7	10.3	8.2	7.25	37.4	366.52

7 Conclusions

In this paper, we study the multi-period production/inventory problem with unobserved lost sales and stockout-based competition. The aggregate demand distribution form in each period is known but one parameter is unknown before selling season. The competing manufacturers update the estimation of demand distribution using the historical sales data. We model the problem as a Bayesian inventory game with three-dimension state-space. We give the conditions under which the dimension of the state-space reduces to two and prove that the game has a unique Nash equilibrium. We also find that the equilibrium solution is decreasing in shape parameter and the equilibrium payoffs of the manufacturers are increasing in shape parameter. We show that the equilibrium strategy is the produce-up-to policy.

1. We assume that all one's unsatisfied customers switch to buy the rival's product if available. In practice, when customers encounter a stockout, only part of the unsatisfied customers looking for substitute product. In such case, how the substitute rate impacts on the equilibrium strategy and payoffs is an interesting topic.
2. We assume two manufacturers are symmetric in cost and revenue information. A natural extension of the Bayesian inventory duopoly is incorporating information asymmetry into analysis. Asymmetric information brings a new dimension to the competitive problem and it is another important topic need to be studied in further.
3. New business models and technology are bringing new challenges to firms. For example, huge volumes of data from e.g. web station, social APP or bricks-and-mortar stores provide chance for understanding competition better and managing inventory effectively. Then incorporating data driven into inventory competition is a challenge problem for firms.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under grant numbers 71962026, the Humanities and Social Sciences Foundation of Ministry of Education of China under grant number 19YJA630111, Natural Science Foundation of Inner Mongolia under grant number 2019MS07004, and Beijing Logistics Informatics Research Base.

References

1. Anderson Consulting, Where to look for incremental sales gains: The retail problem of out-of-stock merchandise. The Coca-Cola Retailing Research Council, Atlanta, 1996
2. J.C. Andraski, J. Haedicke, CPFR: time for the breakthrough? Supply Chain Manage. Rev. **7**(3), 54–60 (2003)
3. Z.M. Avsar, M. Baykal-Gürsoy, Inventory control under substitutable demand: a stochastic game application. Naval Res. Logist. **49**(4), 359–375 (2002)
4. K.S. Azoury, Bayes solution to dynamic inventory models under unknown demand distribution. Manage. Sci. **31**(9), 1150–1160 (1985)

5. A. Bisi, M. Dada, Dynamic learning, pricing, and ordering by a censored newsvendor. Naval Res. Logist. **54**(4), 448–461 (2007)
6. A. Bisi, M. Dada, S. Tokdar, A censored-data multiperiod inventory problem with newsvendor demand distributions. Manuf. Serv. Oper. Manage. **13**(4), 525–533 (2011)
7. D.J. Braden, M. Freimer, Informational dynamics of censored observations. Manage. Sci. **37**(11), 1390–1404 (1991)
8. B. Chen, X. Chao, Dynamic inventory control with stockout substitution and demand learning. Manage. Sci. **66**(3), 4921–5484 (2020)
9. X. Ding, M.L. Puterman, A. Bisi, The censored newsvendor and the optimal acquisition of information. Oper. Res. **50**(3), 517–527 (2002)
10. E.C.R. Europe, *Optimal Shelf Availability: Increasing Shopper Satisfaction at the Moment of Truth* (ECR Europe and Roland Berger, Kontich, 2003).
11. M.A. Lariviere, E.L. Porteus, Stalking information: Bayesian inventory management with unobserved lost sales. Manage. Sci. **45**(3), 346–363 (1999)
12. S.A. Lippman, K.F. Mccardle, The competitive newsboy. Oper. Res. **45**(1), 54–65 (1997)
13. A.J. Mersereau, Demand estimation from censored observations with inventory record inaccuracy. Manuf. Serv. Oper. Manage. **17**(3), 335–349 (2015)
14. M. Nagarajan, R.S. Rajagopalan, A multi-period model of inventory competition. Oper. Res. **57**(3), 785–790 (2009)
15. S. Netessine, N. Rudi, Y. Wang, Inventory competition and incentives to back-order. IIE Trans. **38**(11), 883–902 (2006)
16. T.L. Olsen, R.P. Parker, On Markov equilibria in dynamic inventory competition. Oper. Res. **62**(2), 332–344 (2014)
17. P.B. Schary, M. Christopher, The anatomy of a stockout. J. Retail. **55**(2), 59–70 (1979)
18. J. Zhang, S. He, J. Zhang, T.C.E. Cheng, Purchase and retrieval competition for seasonal produce. Naval Res. Logist. **67**(3), 161–184 (2020)
19. H. Scarf, Bayes solutions of the statistical inventory problem. The Annal. Math. Stat. **30**(2), 490–508 (1959)