Springer Series in Reliability Engineering

Hoang Pham *Editor*

Reliability and Statistical Computing

Modeling, Methods and Applications

Springer

Springer Series in Reliability Engineering

Series Editor

Hoang Pham, Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ, USA

Today's modern systems have become increasingly complex to design and build, while the demand for reliability and cost effective development continues. Reliability is one of the most important attributes in all these systems, including aerospace applications, real-time control, medical applications, defense systems, human decision-making, and home-security products. Growing international competition has increased the need for all designers, managers, practitioners, scientists and engineers to ensure a level of reliability of their product before release at the lowest cost. The interest in reliability has been growing in recent years and this trend will continue during the next decade and beyond.

The Springer Series in Reliability Engineering publishes books, monographs and edited volumes in important areas of current theoretical research development in reliability and in areas that attempt to bridge the gap between theory and application in areas of interest to practitioners in industry, laboratories, business, and government.

Indexed in Scopus

Interested authors should contact the series editor, Hoang Pham, Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA. Email: hopham@rci.rutgers.edu, or Anthony Doyle, Executive Editor, Springer, London. Email: anthony.doyle@springer.com.

More information about this series at http://www.springer.com/series/6917

Hoang Pham
Editor

Reliability and Statistical Computing

Modeling, Methods and Applications

Springer

Editor
Hoang Pham
Department of Industrial and Systems
Engineering
Rutgers University
Piscataway, NJ, USA

ISSN 1614-7839 ISSN 2196-999X (electronic)
Springer Series in Reliability Engineering
ISBN 978-3-030-43411-3 ISBN 978-3-030-43412-0 (eBook)
https://doi.org/10.1007/978-3-030-43412-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

We're living in an era of fast and unpredictable change. Billions of people are connected to each other through their mobile devices and the Internet of Things (IoT). Data is being collected and processed like never before. The era of AI through reliability and statistical machine computing as well as intelligent and recommender systems with almost all applications and service industry has experienced a dramatic shift in the past two decades to a truly global industry, known as the *Industry 4.0*. The forces that have driven this change are still at play and will continue. Most of the products which affect our daily lives are becoming even more complex than ever. This book, consisting of 18 chapters, aims to address both research and practical aspects in reliability and statistical computing with emphasis on the applications. Each chapter is written by active researchers and experienced practitioners in the field aiming to connect the gap between the theory and practice and to trigger new research challenges in reliability and statistical computing of recent complex products and the customer needs in practices.

In chapter "Fatigue Life Distribution Estimation", it discusses an empirically based methodology for estimating the cumulative distribution functions for fatigue life that incorporates available fatigue life data for various stresses given the applied load. Chapter "Reliability Improvement Analysis Using Fractional Failure" describes an approach to improve product reliability during development by using fractional failure analysis method incorporating failure fix effectiveness during each testing and failure fix phase. The chapter also describes a new product introduction reliability improvement process with fractional failure analysis. Chapter "Modelling Innovation Paths of European Firms Using Fuzzy Balanced Scorecard" discusses an approach for modeling innovation paths of European firms and its strategic performance measurement by integrating fuzzy balanced scorecard. The chapter also provides empirical evidence for the effectiveness of the approach based on a large dataset of European firms. Chapter "Innovation Environment in Europe—Efficiency Analysis Case Study" discusses ways to assess the effectiveness, performance, and productivity of comparable production units within EU28 countries based on their innovation performance measured by European Commissions' European Innovation Scoreboard 2017.

In chapter "Application of Artificial Intelligence in Modeling a Textile Finishing Process", it discusses an application of three machine learning techniques such as extreme learning machine, support vector regression, and random forest in modeling a textile finishing process to predict the color properties of ozonated reactive-dyed cottons. Chapter "Developing Alert Level for Aircraft Components" illustrates the development of an alert level for a helicopter air-conditioning system (ACS) incorporating the temperature, humidity, and airflow inside the helicopter for the purpose of improving operational performance.

Reliable network is an important problem in many real-world applications such as traffic planning, computer network planning, and power transmission planning. Chapter "Computation in Network Reliability" discusses procedures and algorithms to compute the reliability of complex networks. The chapter also presents some numerical calculations of some networks to illustrate the proposed algorithms. Chapter "Integrating Sentiment Analysis in Recommender Systems" discusses the use of hybrid deep learning approach for sentiment analysis task and matrix factorization for recommender systems. The chapter also presents an experiment based on Amazon food reviews and its performance of a recommendation system based on root mean square errors score. Chapter "Crowdsourcing Platform for Collecting Cognitive Feedbacks from Users: A Case Study on Movie Recommender System" presents a crowd sourcing-based recommendation platform that can collect and share cognitive feedbacks from users for a case study on movie recommender system.

Aiming to create a safety culture that would help to reduce the driver casualty involved in traffic accidents, chapter "A DD-SHELL HF Model for Bus Accidents" presents an analysis about the causes of the accident with respect to both human factors and organizational lapses using the combination of the aviation dirty dozen factor approach and the SHELL model. Chapter "Development of MI-ANFIS-BBO Model for Forecasting Crude Oil Price" studies the adaptive neuro-fuzzy inference system (ANFIS) with consideration of parameter optimization using the biogeography-based optimization (BBO) algorithm and mutual information (MI) technique for forecasting crude oil price. The approach combines the strengths of fuzzy logic, neural network, and the heuristic algorithm to detect the trends and patterns in crude oil price field data and improve the quality of forecasting. Chapter "Obtaining More Specific Topics and Detecting Weak Signals by Topic Word Selection" presents a method to automatically evaluate specific topics with respect to their usefulness for applications like the detection of new innovations and detect weak signals produced by latent Dirichlet allocation. The chapter also discusses the quality of a topic in terms of both coherence and specificity by using context information from the document corpus.

In chapter "An Improved Ensemble Machine Learning Algorithm for Wearable Sensor Data Based Human Activity Recognition", it presents an improved machine learning method for wearable sensor data based human activity recognition (HAR) using convolution neural networks (CNN) and long short-term memory (LSTM) networks. The proposed method shows an improved performance of these machine learning methods for HAR. Chapter "Average Failure Rate and Its Applications of Preventive Replacement Policies" discusses the average failure rate, which is based

on the conditional failure probability and the mean time to failure, given that the unit is still survival at the mission arrival time where the mission arrival time follows a gamma distribution and the failure time of the unit has a Weibull distribution. The chapter also presents various optimization maintenance models including age replacement policies and the periodic policies with minimal repairs. Chapter "Optimal Maintenance Models of Social Infrastructures Considering Natural Disasters" discusses several optimal maintenance models of social infrastructures considering delay, multiple degradation levels, modified costs which depend on damage levels, and disaster recovery which minimizes the expected cost rates. The cost factors considered in the model include maintenance periods, maintenance time delay, wide variety of preventive maintenance costs and degradation levels, and natural disaster distribution.

In chapter "Optimal Checkpoint Intervals, Schemes and Structures for Computing Modules", it studies a high-reliability computing system with redundancy techniques and recovery methods to prevent failure occurrences considering two checkpoint schemes in which their respective interval times are periodic and random. The chapter discusses the optimal checkpoint interval policies and structures with computing modules for periodic and random checkpoint models. Chapter "Data Envelopment Analysis as a Tool to Evaluate Marketing Policy Reliability" describes the data envelopment analysis design and its applications for effectiveness evaluation of company marketing strategies. The data envelopment analysis approach is a useful tool that can provide suggestions on the optimal marketing strategies to achieve superior performance. Obtaining efficient and reliable software under resource allocation constraint is an important goal of software management science. Chapter "Computational Intelligence Approaches for Software Quality Improvement" discusses the usage of artificial immune algorithms for software testing and recent development paradigms of computational intelligence for software quality improvement.

Postgraduates, researchers, data scientists, and engineers will definitely gain great knowledge with the help of this book in the areas of reliability engineering and safety, applied statistics, machine learning and recommender systems, and its applications. The material is proposed for graduate and advanced undergraduate level students.

I acknowledge Springer for this opportunity and professional support. Importantly, I would like to thank all the chapter authors and reviewers for their availability for this work.

Piscataway, NJ, USA Hoang Pham
January 2020

Contents

Editor and Contributors

About the Editor

Dr. Hoang Pham is a Distinguished Professor and former Chairman (2007–2013) of the Department of Industrial and Systems Engineering at Rutgers University, New Jersey. Before joining Rutgers, he was a Senior Engineering Specialist with the Boeing Company and the Idaho National Engineering Laboratory. He has been served as Editor-in-Chief, Editor, Associate Editor, Guest Editor, and board member of many journals. He is the Editor of *Springer Book Series in Reliability Engineering,* the Editor of *World Scientific Book Series on Industrial and Systems Engineering*, and has served as Conference Chair and Program Chair of over 40 international conferences. He is the author or coauthor of 6 books and has published over 185 journal articles and edited 13 books including *Springer Handbook in Engineering Statistics* and *Handbook in Reliability Engineering.* He has delivered over 40 invited keynote and plenary speeches at many international conferences. His numerous awards include the 2009 IEEE Reliability Society *Engineer of the Year Award.* He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and the Institute of Industrial Engineers (IIE).

Contributors

Grigore Albeanu "Spiru Haret" University, Bucharest, Romania

Jiajia Cai College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Shin-Guang Chen Tungnan University, New Taipei City, Taiwan

Zaytsev Dmitry International Laboratory for Applied Network Research, National Research University Higher School of Economics, Moscow, Russia

Quang Hung Do Faculty of Information Technology, University of Transport Technology, Hanoi, Vietnam

D. Gary Harlow Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA, USA

Michael Grottke GfK SE, Global Data Science, Nürnberg, Germany

Zhenglei He GEMTEX – Laboratoire de Génie et Matériaux Textiles, ENSAIT, Lille, France

Bui Thanh Hung Data Analytics & Artificial Intelligence Laboratory, Faculty of Engineering and Technology, Thu Dau Mot University, Thu Dau Mot, Binh Duong, Vietnam

Petr Hájek Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic

Kodo Ito Department of Social Management Engineering, Graduate School of Engineering, Tottori University, Tottori, Koyama-cho Minami, Japan

Mingxiao Jiang Minneapolis, MN, USA

Jason J. Jung Department of Computer Engineering, Chung-Ang University, Seoul, Korea

Ludovic Koehl Ecole Nationale Supérieure des Arts et Industries Textiles, GEMTEX Laboratory, Roubaix, France

Takumi Kishida Department of Social Management Engineering, Graduate School of Engineering, Tottori University, Tottori, Koyama-cho Minami, Japan

Michaela Kotková Stříteská Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic

Laura Kölbl Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany

Henrik Madsen Danish Technical University, Lyngby, Denmark

Wai Yeung Man Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Satoshi Mizutani Department of Business Administration, Aichi Institute of Technology, Toyota, Japan

Toshio Nakagawa Department of Business Administration, Aichi Institute of Technology, Toyota, Aichi, Japan

Kenichiro Naruse Nagoya Sangyo University, Yamanota Araicho, Owariasahi, Aichi, Japan

Huu Du Nguyen Division of Artificial Intelligence, Dong-A University, Danang, Vietnam

Luong Vuong Nguyen Department of Computer Engineering, Chung-Ang University, Seoul, Korea

Kelvin K. F. Po Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Florin Popenţiu-Vlădicescu University Politehnica of Bucharest Bucharest, Romania; Academy of Romanian Scientists, Bucharest, Romania

Viktor Prokop Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic

Jan Stejskal Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic

Feng-Bin Sun Tesla, Inc., Palo Alto, CA, USA

Guillaume Tartare Ecole Nationale Supérieure des Arts et Industries Textiles, GEMTEX Laboratory, Roubaix, France

Sébastien Thomassey GEMTEX – Laboratoire de Génie et Matériaux Textiles, ENSAIT, Lille, France

Kim Phuc Tran ENSAIT & GEMTEX, Roubaix, France

Kuskova Valentina International Laboratory for Applied Network Research, National Research University Higher School of Economics, Moscow, Russia

Eric T. T. Wong Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Changhai Yi National Local Joint Engineering Laboratory for Advanced Textile Processing and Clean Production, Wuhan Textile University, Wuhan, China

Higuchi Yoshiyuki Faculty of Symbiotic Systems Science, Fukushima University, Kanayagawa, Fukushima, Japan

Xianyi Zeng ENSAIT & GEMTEX, Roubaix, France

Xufeng Zhao College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Fatigue Life Distribution Estimation

D. Gary Harlow

Abstract Modeling fatigue life is complex whether it is applied to structures or experimental programs. Through the years several empirical approaches have been utilized. Each approach has positive aspects; however, none have been acceptable for every circumstance. On many occasions the primary shortcoming for an empirical method is the lack of a sufficiently robust database for statistical modeling. The modeling is exacerbated for loading near typical operating conditions because the scatter in the fatigue lives is quite large. The scatter may be attributed to microstructure, manufacturing, or experimental inconsistencies, or a combination thereof. Empirical modeling is more challenging for extreme life estimation because those events are rare. The primary purpose herein is to propose an empirically based methodology for estimating the cumulative distribution functions for fatigue life, given the applied load. The methodology incorporates available fatigue life data for various stresses or strains using a statistical transformation to merge all the life data so that distribution estimation is more accurate than traditional approaches. Subsequently, the distribution for the transformed and merged data is converted using change-of-variables to estimate the distribution for each applied load. To assess the validity of the proposed methodology percentile bounds are estimated for the life data. The development of the methodology and its subsequent validation is illustrated using three sets of fatigue life data which are readily available in the open literature.

Keywords Cycle dependent parameters · Data fusion · Fatigue life transformation · Weibull distribution

Nomenclature

α, β, γ	Weibull cdf parameters
α_s	significance

D. Gary Harlow (✉)
Mechanical Engineering and Mechanics, Lehigh University, 19 Memorial Drive West, Bethlehem, PA 18015–3085, USA
e-mail: dgh0@lehigh.edu

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_1

AD	Anderson–Darling goodness of fit test
cdf	cumulative distribution function
cv	coefficient of variation
$\Delta\varepsilon$	strain range
$\Delta\sigma$	stress range
FLT	Fatigue Life Transformation
$F(\bullet)$	cdf
KS	Kolmogorov–Smirnov goodness of fit test
MLE	maximum likelihood estimation
N_A	arbitrary normalization constant
N_f	cycles to failure
n	sample size
m	number of different values of applied stress
p	percentile
s	sample standard deviation
s_A	arbitrary normalization constant
S–N	stress–number of cycles
t	time, cycles
$W(\alpha, \beta, \gamma)$	three parameter Weibull cdf
$y_{1/2}$	median
$\bar{y}$	sample average

1 Introduction

A difficulty with fatigue life data is the characterization of its variability, which can be several orders of magnitude [1], especially for loading near operating conditions. The variability is attributable to experimental error, as well as material microstructure or processing. Thus, estimation and prediction of fatigue life is challenging. A key concern is the characterization of the cumulative distribution function (cdf) for fatigue life, given an applied load, which may be either the stress range $\Delta\sigma$ or the strain range $\Delta\varepsilon$. The lower tail portion of the cdf which depicts high reliability is especially critical; however, that is where scatter is more pronounced, and sample sizes are smaller. Empirically modeling fatigue life has been considered numerous times. A simple internet search for *statistical fatigue life modeling* yields in excess of 10 million citations. A relatively recent work is [2], in which the authors incorporate statistical analysis with traditional stress-life, strain-life, or crack propagation models. While empiricism is used, the thrust is to incorporate as much physically motivated modeling as possible. More frequently, investigators attempt to fit a stress-life (S-N) curve through the data, especially the medians for given $\Delta\sigma$ or $\Delta\varepsilon$. A nice review of such practices is [3]. Another example of statistical modeling of fatigue data is contained in [4] in which over two chapters are devoted statistical methodologies. Other examples of statistical stress-life analysis may be found in [5, 6]. While there

are many more papers, books, and conferences publications on statistical fatigue modeling, these are representative. In spite of all these references, statistical fatigue analysis is still an open area of investigation.

Herein, an empirically based approach for accurately estimating the fatigue life cdf, given $\Delta\sigma$ or $\Delta\varepsilon$ is proposed. The methodology merges fatigue life data using a statistical transformation for the estimation. The statistical technique increases the sample size by merging fatigue data for more precise assessment. This is necessitated because there is often large variability in S-N data, and the sample sizes are small. Validation of the modeling is essential, especially for prediction of life outside of the range of experimental observations. The validity of the methodology is evaluated by considering percentile bounds estimated for the S-N data. The development of the methodology and its subsequent validation is illustrated using three different fatigue life datasets.

A fundamental issue in fatigue life estimation is the choice of an underlying cdf. The cdf used in the ensuing analyses is a three-parameter Weibull cdf. A fairly recent example of a traditional statistical S-N methodology using the three-parameter Weibull cdf is [7] where fatigue of structural and rolling contact problems are considered. There seems to be a need for more experimental data to enhance modeling in almost all fatigue analyses. This is addressed in [8] by using normalization for the fatigue life data so that all the data are merged. The normalized data are then modeled with a three-parameter Weibull cdf. Even though the intention in [8] is similar to the emphasis herein, the methodology is somewhat different.

2 Fatigue Life Data

Fatigue life data are most often presented on an S-N plot which shows the fatigue data for a given load. The load is typically stress or strain. Thus S-N can represent stress-life or strain-life. An additional way in which the fatigue data are presented is on a probability plot. Both of these representations will be used subsequently. Three different sets of fatigue life data are considered for the proposed method.

The first set considered is one of the special cases taken from [9]. Fatigue testing was conducted on 2024-T4 aluminum alloy specimens. The fatigue tests were performed on rectangular specimens with dimensions of 110 mm long, 52 mm wide, and 1 mm thick, and with a center cut circular hole of radius 5 mm. Holes were cut using standard procedures with a lathe, and burrs were removed by polishing techniques. Testing was conducted in laboratory environment where temperatures of 295–297 K (approximately 22–24 °C) and relative humidities of 50–56% were observed. Constant amplitude tests were performed at a frequency of 30 Hz on a single machine with a single operator in order to minimize experimental error. Because of the extensiveness of this data, they have been used in a variety of analyses; e.g., [10–12]. These data are summarized in Table 1. A total of 222 specimens were tested using eight different values for $\Delta\sigma$. The specimens were tested to fracture. The sample coefficients of variation (cv) are nearly the same, approximately 9%, for the larger

Table 1 Statistical summary of fatigue life data for 2024-T4 specimens [9]

Load, $\Delta\sigma$ (MPa)	Size (m)	Average ($\bar{x}$)	Standard deviation (s)	*cv* (%)
255	21	18,200	1760	9.6
235	30	28,700	2500	8.7
206	30	59,400	4230	7.1
177	30	146,000	12,600	8.6
157	30	264,000	22,600	8.6
137	30	519,000	96,200	18.5
127	30	1,710,000	1,090,000	63.8
123	21	4,530,000	2,660,000	58.7

values of $\Delta\sigma$. When $\Delta\sigma$ is 137 MPa, the *cv* is about double, and for the two smaller values of $\Delta\sigma$ the scatter increases significantly. The fatigue life data are plotted on an S-N graph in Fig. 1 in the traditional linear versus logarithm S-N format. As $\Delta\sigma$ is reduced, the increase in the scatter in life is apparent. Modeling the increasing variability for decreasing $\Delta\sigma$is the challenge for accurate fatigue life prediction.

The second set of fatigue data to be considered is data collected at room temperature for ASTM A969 hot dipped galvanized sheet steel with a gauge thickness of 1.78 mm [13]. ASTM A969 is a cold-rolled, low carbon, extra deep drawing steel. This steel is very ductile and soft, and it is age resistant. The automotive industry uses it in applications where severe forming is required, e.g., inner door components, dash panels, body side components, and floor pans with spare tire tubs. Fatigue tests for the ASTM A969 specimens were conducted using a triangular waveform at 25 Hz. The fatigue tests were terminated, i.e., designated as a failure, when the tensile load dropped by 50% of the maximum load. A total of 69 specimens were tested to failure. The data are summarized in Table 2. The *cv* given $\Delta\varepsilon$ is more scattered than those for the 2024-T4 data. The smallest value of $\Delta\varepsilon$ has the largest scatter, but the second

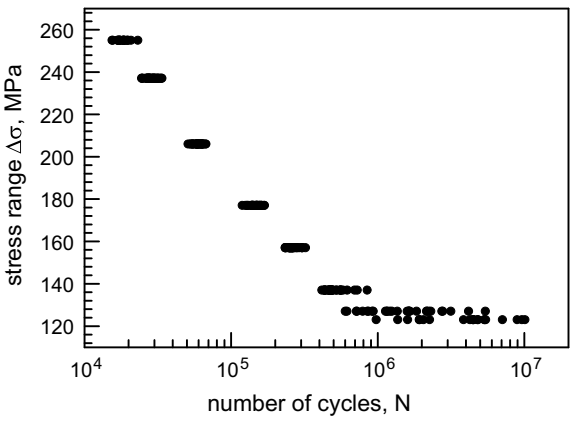


Fig. 1 Fatigue life data for 2024-T4 specimens [9]

Table 2 Statistical summary of fatigue life data for ASTM A969 specimens [13]

Load, $\Delta\varepsilon$ (mm/mm)	Size (m)	Average ($\bar{x}$)	Standard deviation (s)	cv (%)
0.0080	3	4,100	815	19.8
0.0060	12	14,000	3,800	27.2
0.0050	6	34,600	7,030	20.3
0.0040	12	56,500	8,860	15.7
0.0030	12	107,000	12,900	12.1
0.0024	6	199,000	29,800	15.0
0.0020	11	499,000	68,300	13.7
0.0018	7	1,030,000	391,000	37.8

largest $\Delta\varepsilon$ is rather large as well. This behavior is seen graphically on the S-N plot in Fig. 2. Consequently, these data are not as statistically well behaved as the 2024-T4 data. These data have been used to investigate other types of fatigue modeling [14, 15].

The third set to assess is 9Cr-1Mo steel which were collated from a round-robin test program and were reported in [16]. This steel is creep strengthen, and it is frequently used in thermal power plants to improve the energy efficiency of the power plant by increasing operating temperatures and pressures. Specifically, 9Cr-1Mo is often used for steam generator components of both fossil fired and nuclear power plants. The material from which the data were generated was a single cast, rolled plate with a nominal tensile strength of 623 MPa [16]. A total of 130 specimens were tested to failure. The data are summarized in Table 3. The cv given $\Delta\varepsilon$ is even more scattered than the above datasets. In fact, there does not seem to be any discernible pattern. The data are shown on Fig. 3. The reason for the unusual statistical behavior would require more in depth analysis than is provided in [16]. Usually round-robin testing has considerably more variability in results because testing conditions and

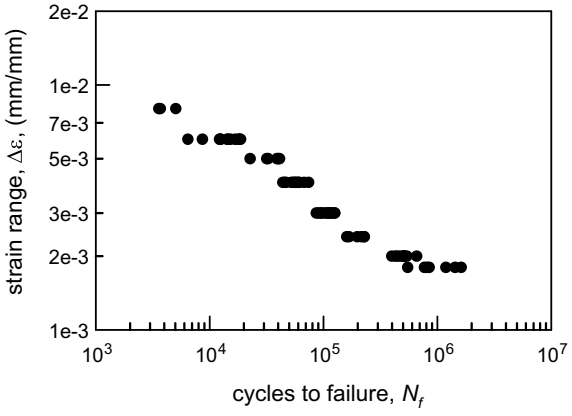


Fig. 2 Fatigue life data for ASTM A969 specimens [13]

Table 3 Statistical summary of fatigue life data for 9Cr-1Mo specimens [16]

Load, $\Delta\varepsilon$ (mm/mm)	Size (m)	Average ($\bar{x}$)	Standard deviation (s)	cv (%)
0.021	5	472	88	18.8
0.020	32	542	189	34.9
0.019	8	572	345	60.2
0.012	34	1,260	357	28.4
0.011	6	1,290	475	36.9
0.006	37	4,820	2,000	41.6
0.005	8	12,700	7,700	60.8

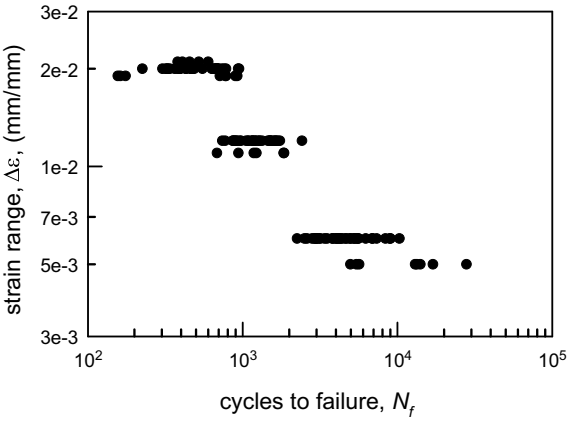


Fig. 3 Fatigue life data for 9Cr-1Mo specimens [16]

methodologies are not consistent. Nevertheless, the data will serve as an excellent case for the proposed modeling approach.

3 Data Fusion for Fatigue Life Analysis

The following is a purely empirical method to improve fatigue life modeling. Because fatigue data are usually limited in number for relatively few different loading conditions, modeling is crude. A methodology that has been developed to account for uncertainty for static properties [17] is adapted for fatigue life data. The basis of the approach is a linear transformation of a collection of experimental observations $\{y_j : 1 \leq j \leq n\}$ into another set of values $\{z_j : 1 \leq j \leq n\}$ so that both sets have the same average and sample standard deviation. Let

$$z_j = ay_j + b. \tag{1}$$

The choices of a and b in Eq. (1) are easily determined by simple algebra to be the following:

$$a = \frac{s_A}{s_y} \text{ and } b = N_A - \frac{s_A}{s_y}\bar{y}, \tag{2}$$

where $\bar{y}$ is the average and s_y is the sample standard deviation for $\{y_j : 1 \leq j \leq n\}$, and N_A and s_A are arbitrary values chosen for normalization.

For fatigue data the life times are distributed over several orders of magnitude that the procedure in Eqs. (1) and (2) is applied to the natural logarithm of the life times. Let m be the number of different values of applied stress or strain, i.e., $\{\Delta\sigma_k : 1 \leq k \leq m\}$ or $\{\Delta\varepsilon_k : 1 \leq k \leq m\}$. Given $\Delta\sigma_k$ or $\Delta\varepsilon_k$ the associated life times are $\{N_{k,j} : 1 \leq j \leq n_k\}$ where n_k is its sample size. Let

$$y_{k,j} = \ln(N_{k,j}) \tag{3}$$

be the transformed life times. Substituting Eq. (2) into Eq. (1) leads to the following:

$$z_{k,j} = \frac{s_A}{s_{y,k}}(y_{k.j} - \bar{y}_k) + N_A. \tag{4}$$

Thus, the averages and sample standard deviations of $\{y_{k,j} : 1 \leq j \leq n_k\}$ and $\{z_{k,j} : 1 \leq j \leq n_k\}$ are equal to each other. The next step is to merge all the transformed $z_{k,j}$ values from Eq. (4) for $1 \leq j \leq n_k$ and $1 \leq k \leq m$. The purpose in using the merged values is to have a more extensive dataset for estimation of the cdf. This is especially critical for estimating the extremes of the cdf more accurately. Subsequently, an appropriate cdf $F_Z(z)$ is found that characterizes the merged data. It is assumed that this cdf also characterizes the subsets $\{z_{k,j} : 1 \leq j \leq n_k\}$ of the merged set. With this assumption and the linear transformation in Eq. (4), the cdfs for $\{y_{k,j} : 1 \leq j \leq n_k\}$ $F_{y,k}(y)$ can be derived from $F_Z(z)$ as follows:

$$F_{y,k}(y) = F_Z(\frac{s_A}{s_{y,k}}(y - \bar{y}_k) + N_A). \tag{5}$$

The approach is designated as the Fatigue Life Transformation (FLT). Recall that the above methodology is applied to natural logarithms. In order to make observations on the actual fatigue lives, the values must be changed back to actual cycles.

4 Flt Analysis for 2024-T4 Fatigue Life Data

To evaluate the effectiveness of the proposed FLT methodology, the fatigue life data summarized in Table 1 and shown on Fig. 1 is considered. Recall that the FLT is

applied to the natural logarithm of the fatigue data; see Eq. (3). The arbitrarily chosen values for N_A and s_A are 26 and 1, respectively. The rather large value for N_A was chosen to assure that $z_{k,j}$ in Eq. (4) is positive. The 222 FLT data are shown on Fig. 4, where the axes are labeled to be easily read. Each set of data for a given $\Delta\sigma_k$ are transformed using FLT. These transformed data are well grouped so that it is reasonable to merge them. Figure 5 shows the entire 222 FLT values merged into a common sample space. The FLT merged data contain approximately 7–10 times more data than those for each given $\Delta\sigma_k$. Thus, estimation for the cdf is necessarily more accurate which results in a better characterization of its lower tail. Also, notice that the cycles are transformed using the FLT procedure; they are not actual cycles to failure, i.e., they are not equivalent to the data shown on Fig. 1. The solid line is the maximum likelihood estimation (MLE) for a three–parameter Weibull cdf W(α, β, γ), where α is the shape parameter, β is the scale parameter, and γ is the location parameter. The form of W(α, β, γ) is

$$F(x) = 1 - \exp\{-[(x-\gamma)/\beta]^{\alpha}\}, \; x \geq \gamma. \tag{6}$$

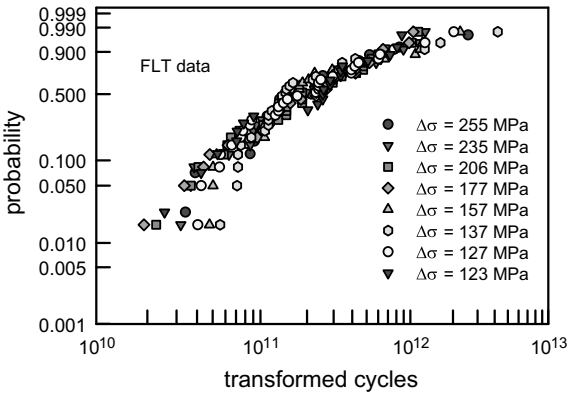


Fig. 4 FLT fatigue life data for 2024-T4 specimens given $\Delta\sigma$ [9]

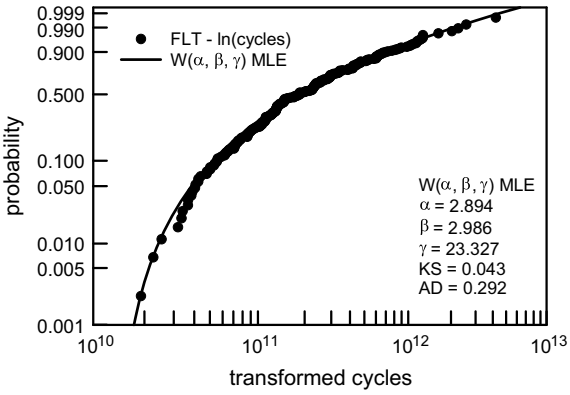


Fig. 5 Merged FLT fatigue life data for 2024-T4 specimens [9]; Weibull MLE

Graphically, the fit is excellent. The Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) goodness of fit test statistics are 0.043 and 0.292, respectively. Both of which indicate that the MLE is acceptable for any significance level α_s less than 0.3. Consequently, the MLE W(α, β, γ) cdf is an excellent representation of the merged FLT data. The MLE estimated parameters are $\hat{\alpha} = 2.894$; $\hat{\beta} = 2.986$; and $\hat{\gamma} = 23.327$.

The three–parameter Weibull cdf W(α, β, γ) shown in Eq. 6 was selected for consideration because it has become a very popular cdf to represent fatigue data since its namesake used it for that purpose [18]. Two popular resources for the Weibull cdf, which advocate its use and contain examples of its applications, are [19, 20]. The primary reason for its choice, however, is because there is a minimum value represented by γ. Typically, any collection of fatigue life data is spread over two or three orders of magnitude. Consequently, a nonzero minimum value is required to appropriately represent the fatigue data.

Now, it is assumed that the MLE estimated parameters for the FLT merged data are acceptable for each of its subsets $\{z_{k,j} : 1 \le j \le n_k\}$. The cdf $F_{y,k}(y)$ for each given $\Delta\sigma_k$ can be determined using Eq. (5) as follows:

$$F_{y,k}(y) = 1 - \exp\{-[(\frac{s_A}{s_{y,k}}(y - \bar{y}_k) + N_A - \hat{\gamma})/\hat{\beta}]^{\hat{\alpha}}\}. \tag{7}$$

Recall that the arbitrary constants N_A and s_A are 26 and 1, respectively. Equation (7) can be rewritten to put it into the standard Weibull cdf form W(α, β, γ);

$$F_{y,k}(y) = 1 - \exp\{-[(y - [\bar{y}_k + (\hat{\gamma} - N_A)(\frac{s_{y,k}}{s_A})])/\hat{\beta}(\frac{s_{y,k}}{s_A})]^{\hat{\alpha}}\}, \tag{8}$$

where the shape parameter $\hat{\alpha}$ is the same for each individual cdf, but the scale parameter $\hat{\beta}_k$ and location parameter $\hat{\gamma}_k$ are

$$\hat{\beta}_k = \hat{\beta}(\frac{s_{y,k}}{s_A}); \text{ and } \hat{\gamma}_k = \bar{y}_k + (\hat{\gamma} - N_A)(\frac{s_{y,k}}{s_A}), \tag{9}$$

which are explicitly dependent on the sample parameters for the fatigue life data given $\Delta\sigma_k$ and the arbitrary constants $N_A$and s_A. Recall that the FLT is for $\ln(N_f)$; consequently, the range of values is considerably smaller than that for N_f. Table 4 contains the FLT cdf W(α, β_k, γ_k) parameters for each given $\Delta\sigma_k$. Figure 6 shows the fatigue life data plotted on two parameter Weibull probability paper with the corresponding FLT cdfs W(α, β_k, γ_k). Graphically, these cdfs appear to fit the data well. Indeed, the KS goodness of fit test indicates that all these cdfs are acceptable for any α_s less than 0.25. The AD test, which focuses on the quality of the fit in the tails, is more discriminating. The cdfs when $\Delta\sigma_k$ equals 127, 177, 206, 235, or 255 MPa are acceptable for any α_s less than 0.25. When $\Delta\sigma_k$ equals 123 or 157 MPa, however, the cdfs are acceptable for any α_s less than 0.05. When $\Delta\sigma_k$ equals 137, the AD test implies that the cdf is not acceptable. Although it is not obvious on Fig. 6

Table 4 FLT Weibull parameters for natural logarithm of fatigue life data for 2024-T4 specimens [9]

Load, $\Delta\sigma$ (MPa)	$\hat{\alpha}$	$\hat{\beta}_k$	$\hat{\gamma}_k$
255	2.894	0.280	9.556
235	2.894	0.258	10.032
206	2.894	0.214	10.799
177	2.894	0.261	11.657
157	2.894	0.247	12.259
137	2.894	0.491	12.705
127	2.894	1.678	12.690
123	2.894	1.972	13.375

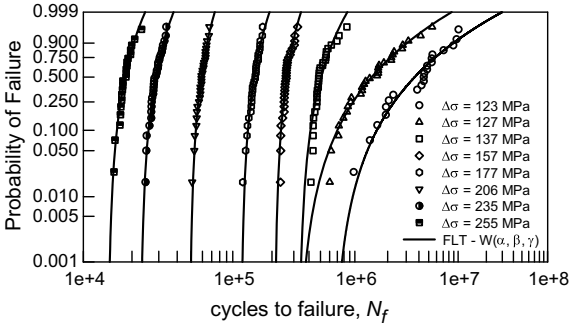


Fig. 6 FLT Weibull cdfs for fatigue life data for 2024-T4 specimens [9]

because the cycles scale is so large, both the upper and lower tail of the FLT cdf are sufficiently different from the life data. Even so, the overall deduction is that the FLT transformation is acceptable for these fatigue life data. Again, the KS test supports that conclusion, and there is only one value for $\Delta\sigma_k$, 137 MPa, for which the AD test indicates otherwise.

Another way to assess the quality of the proposed FLT methodology is to consider percentiles p of the estimated cdfs. The percentiles are given by

$$y_p = \hat{\gamma}_k + \hat{\beta}_k[-\ln(1-p)]^{1/\hat{\alpha}}, \tag{10}$$

which are computed from Eqs. (8) and (9). One of the most common percentiles that is considered is the median $y_{1/2}$, i.e., p is 0.5. Figure 7 is an S-N graph, identical to Fig. 1, where the solid line is the FLT estimated median, and the dashed lines are the FLT estimated 99% percentile bounds. The 99% bounds are very tight while encapsulating the entire set of data for each $\Delta\sigma_k$. They also follow the trend in the S–N data in that they are very narrow when the life data have very little variability, but they are broader when the life data have larger variability. This lends credence to the FLT approach for the 2024-T4 fatigue data.

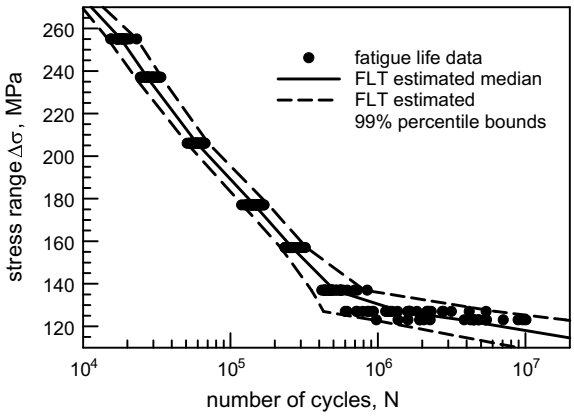


Fig. 7 Fatigue life data for 2024-T4 specimens [9] with FLT percentile bounds

5 FLT Analysis for ASTM A969 Fatigue Life Data

The second applications of the FLT method is for the ASTM A969 fatigue data. Again, the arbitrarily chosen values for N_A and s_A are 26 and 1, respectively. Figure 8 shows the entire 69 FLT values merged into a common sample space. The solid line is the MLE W(α, β, γ), Eq. (6). The KS and AD goodness of fit test statistics are 0.047 and 0.524, respectively. The MLE is acceptable according to the KS test for any significance level α_s less than 0.3. For the AD test, however, it is acceptable only for α_s less than 0.2 because there is some deviation between the data and the MLE in the lower tail. Even so, the FLT data are well represented by the MLE W(α, β, γ) cdf. The MLE estimated parameters are $\hat{\alpha} = 3.057$; $\hat{\beta} = 2.964$; and $\hat{\gamma} = 23.289$.

Using Eqs. (7)–(9), Fig. 9 has the ASTM A969 data with the FLT cdfs W(α, β_k, γ_k). Graphically, the FLT cdfs appear to characterize the data well. In fact, the KS test indicates that all these cdfs are acceptable for any α_s less than 0.3. The AD test, however, implies that the FLT cdfs are marginal, at best. Clearly, the FLT cdfs are

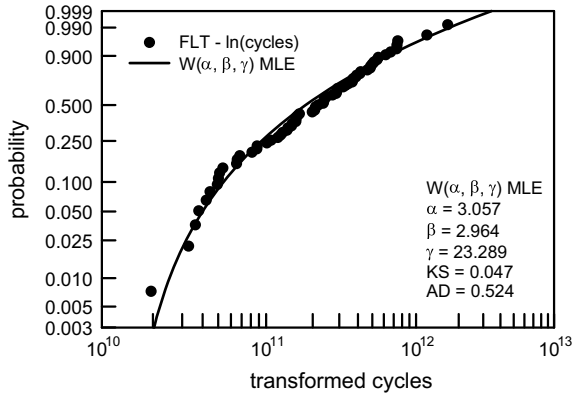


Fig. 8 Merged FLT fatigue life data for ASTM A969 specimens [13]; Weibull MLE

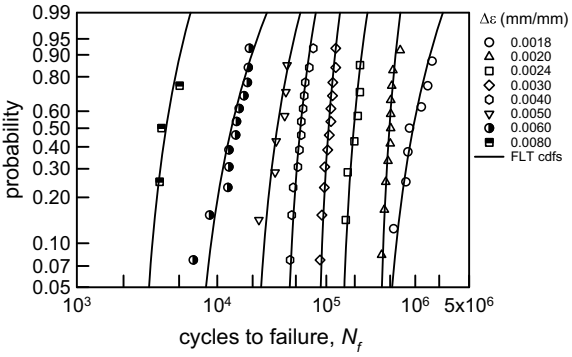


Fig. 9 FLT Weibull cdfs for fatigue life data for ASTM A969 specimens [13]

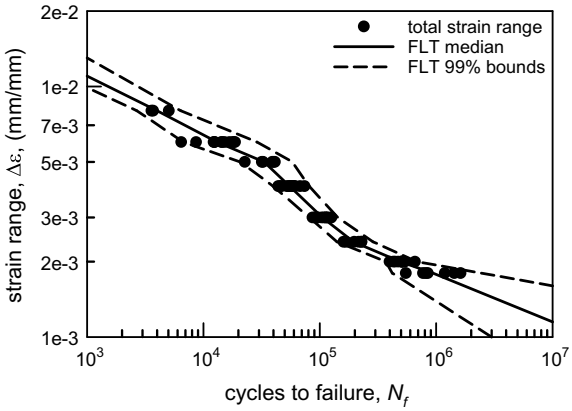


Fig. 10 Fatigue life data for ASTM A969 specimens [13] with FLT percentile bounds

not as accurate in the tails. No doubt, larger samples for each $\Delta\varepsilon_k$ would help with characterization of the extremes. Using Eq. (10), Fig. 10 is an S-N graph, identical to Fig. 2, with the addition of the FLT estimated median, and the FLT estimated 99% percentile bounds. The 99% bounds are very tight, and the all the data are within the bounds for each $\Delta\varepsilon_k$. They are somewhat jagged because they follow the pattern of the S-N data. The analysis is not as crisp as that for the 2024-T4 data; however, there seems to be merit in using the FLT approach for the ASTM A969 fatigue data.

6 FLT Analysis for 9Cr-1Mo Fatigue Life Data

The third application of the FLT method for the 9Cr-1Mo fatigue data does not perform very well. Figure 3 shows unusually large scatter for the higher values of $\Delta\varepsilon_k$ which is a good test for the FLT methodology. The arbitrary scaling factors are the same; N_A and s_A are 26 and 1, respectively. Figure 11 shows the 130 FLT merged values with the MLE $W(\alpha, \beta, \gamma)$. The KS and AD goodness of fit test statistics

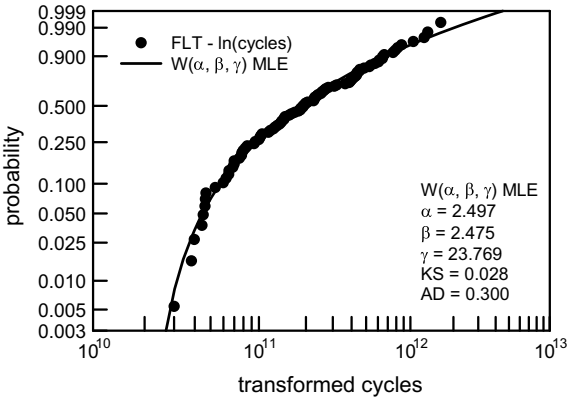


Fig. 11 Merged FLT fatigue life data for 9Cr-1Mo specimens [16]; Weibull MLE

are 0.028 and 0.300, respectively. The MLE is acceptable according to the KS and AD tests for any significance level α_s less than 0.3. The merged data are very well characterized by this cdf. The corresponding MLE estimated parameters are $\hat{\alpha} = 2.497;\ \hat{\beta} = 2.475;$ and $\hat{\gamma} = 23.769$.

The FLT cdfs W(α, β_k, γ_k), Eqs. (7)–(9), for 9Cr-1Mo are shown on Fig. 12. Graphically, the FLT cdfs appear to be acceptable, at least for the cases with more data. Clearly, when $\Delta\varepsilon_k$ is 0.019, the fit is borderline. The KS test indicates that all these cdfs are acceptable for any α_s less than 0.3. On the other hand, the AD test indicates that none of the FLT cdfs are acceptable. The compressed graphical scale for the cycles to failure masks the poor fit of the FLT cdfs to the tails of the data. Figure 13 is the S-N graph, Fig. 3, with the FLT estimated median, and the FLT estimated 99% percentile bounds. Because the FLT cdfs are not very representative of the fatigue data, the 99% bounds are erratic. All the data are within the bounds for each $\Delta\varepsilon_k$. That is somewhat positive. They are quite jagged because they follow the S-N data pattern. One of the difficulties in modeling these data are that there are three values for $\Delta\varepsilon_k$ that have several replicates, but the other values have only a few. Also the scatter in the data when $\Delta\varepsilon_k$ is 0.020 appears to be larger than expected.

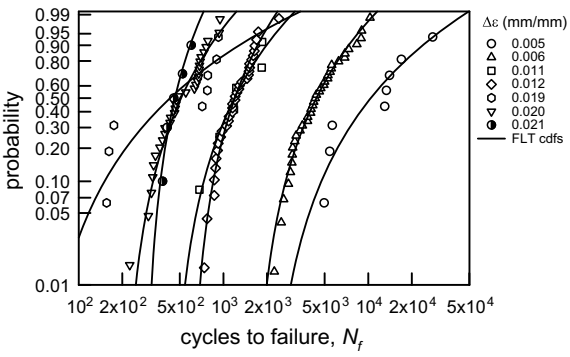


Fig. 12 FLT Weibull cdfs for fatigue life data for 9Cr-1Mo specimens [16]

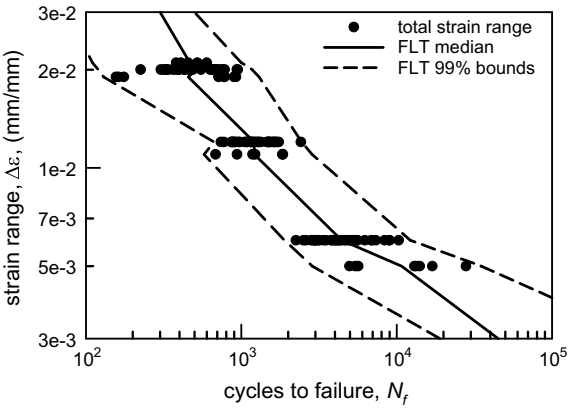


Fig. 13 Fatigue life data for 9Cr-1Mo specimens [16] with FLT percentile bounds

Since there is no explanation for this in [16], the reason would purely speculation. The analysis is marginal for the 9Cr-1Mo data. There may be merit in using the FLT approach for some insight, but conclusions would need to be made very cautiously.

7 Observations and Conclusions

Three sets of fatigue life data were considered for the proposed FLT method. These datasets were selected because of they have replicate data for several applied stress or strain ranges, $\Delta\sigma_k$ or $\Delta\varepsilon_k$, respectively. The primary reason for the proposed FLT method is to accurately model the statistical nature of fatigue life data. Specifically, the estimation of underlying cdfs is crucial. It is well known that fatigue life data have rather large amounts of variability particularly for applied loads similar to typical operating conditions. Modeling these data is very challenging. Associated with this is that many fatigue life data sets have relatively few choices for $\Delta\sigma_k$ or $\Delta\varepsilon_k$, and each choice has limited observations.

The FLT introduced in this paper attempts to help improve fatigue life modeling by using a statistically based transformation to merge data, thereby increasing the effective sample size. The FLT approach transforms the fatigue life data for each given $\Delta\sigma_k$ or $\Delta\varepsilon_k$ so that the averages and standard deviations are the same. Subsequently, the data are merged, and a suitable cdf is statistically estimated for the entire collection. The cdf for each given $\Delta\sigma_k$ or $\Delta\varepsilon_k$ is obtained by standard change-of-variables methods using the cdf that characterizes the entire transformed and merged data.

Using the 2024-T4 fatigue life in [9], the FLT is very promising partly because there are eight different values for $\Delta\sigma_k$ and a total of 222 data being considered. A three–parameter Weibull cdf $W(\alpha, \beta, \gamma)$ is an excellent representation of the merged FLT data, and a $W(\alpha, \beta, \gamma)$ is also appropriate characterization for the underlying cdfs given $\Delta\sigma_k$. Consequently, there is assurance that the lower tail behavior is adequately

modeled because of the methodology. Additionally, the validation for the approach is the computed FLT percentiles for the S-N data. The computed FLT 99% percentiles fully encompass all the S-N data, but more importantly, the bounds are quite tight. The conclusion from this analysis is that the FLT methodology is warranted.

To corroborate this conclusion, two other sets of fatigue life data were considered; ASTM A969 [13] and 9Cr-1Mo [16]. In both cases, the merged FLT data are well characterized by a $W(\alpha, \beta, \gamma)$. The KS test indicates that the transformed $W(\alpha, \beta, \gamma)$ is also suitable for the underlying cdfs given $\Delta\varepsilon_k$, but the AD test is more discriminating. The tail behavior of the underlying cdfs given $\Delta\varepsilon_k$ are marginally acceptable at best, if at all. The 99% percentiles for the ASTM A969 S-N data are quite good. They are encompass the data, and they are tight. For the 9Cr-1Mo S-N data, however, the percentile lines are not very regular. They do encompass the data, but the data have so much variability that little is gained by the analysis.

Based on these three examples, the FLT approach should be employed for fatigue life data analysis when an empirical method is desired. The FLT method excellent for one of the cases, 2024-T4, acceptable for another case, ASTM A969, and marginal for the other case, 9Cr-1Mo. As with all empirical analyses, caution must be exercised when it is implemented. Limited applied loads with limited replicate data for each load hinders accurate modeling for any method including the FLT. As with all empirical methods, the more data there is, the better the accuracy will be. The example which was the worst, 9Cr-1Mo, seems to be poor because there is overly large scatter in the data for the higher loading conditions coupled with applied loads with only a few replicates. Again, the FLT methodology should be implemented with care.

Many sets of experimental fatigue life data contain censoring. This will be investigated in the future. In this case the cdf estimation is more advanced, especially for a three-parameter cdf. In principle the FLT methodology should be similar except for the adjustment for censoring. All things considered, the proposed FLT approach has sufficient promise that further investigation and analysis is certainly warranted. The overarching observation is that the FLT approach is useful if the fatigue data are reasonably well behaved.

References

1. Harlow DG, Wei RP, Sakai T, Oguma N (2006) Crack growth based probability modeling of S-N response for high strength steel. Int J Fatigue 28:1479–1485
2. Castillo E, Fernández-Canteli A (2009) A unified statistical methodology for modeling fatigue damage. Springer Science and Business Media B.V., Berlin
3. Schneider CRA, Maddox SJ (2003) Best practice guide on statistical analysis of fatigue data. International Institute of Welding, IIW-XIII-WG1-114-03, United Kingdom
4. Collins JA (1993) Failure of materials in mechanical design: analysis, prediction, prevention, 2nd edn. Wiley-Interscience Publications, New York
5. Little RE, Ekvall JC (eds) (1981) Statistical analysis of fatigue data. ASTM STP744, American Society for Testing and Materials, Philadelphia

6. Zheng X-L, Jiang B, Lü H (1995) Determination of probability distribution of fatigue strength and expressions of P–S–N curves. Eng Fracture Mech 50:483–491
7. Shimizu S (2005) P–S–N/P–F–L curve approach using three-parameter Weibull distribution for life and fatigue analysis of structural and rolling contact components. Tribol Trans 48:576–582
8. Fernández–Canteli A, Blasón S, Correia JAFO, de Jesus AMP (2014) A probabilistic interpretation of the miner number for fatigue life prediction. FratturaedIntegritàStrutturale 30:327–339
9. Shimokawa T, Hamaguchi Y (1985) Relationship between fatigue life distribution, notch configuration, and S-N curve of a 2024–T4 Aluminum Alloy. Trans ASME J Eng Mater Technol 107:214–220
10. Harlow DG (2005) Probability versus statistical modeling: examples from fatigue life prediction. Int J Reliab Qual Saf Eng 12:1–16
11. Harlow DG (2011) Generalized probability distributions for accelerated life modeling. SAE Int J Mater Manuf 4:980–991
12. Sakai T, Nakajima M, Tokaji K, Hasegawa N (1997) Statistical distribution patterns in mechanical and fatigue properties of metallic materials. Mater Sci Res Int 3:63–74
13. IF_DDQ_HDG70G_Strain_Life_Fatigue.xls www.autosteel.org (2004). Automotive Applications Council, Steel Market Development Institute, Auto/Steel–Partnership Fatigue Data
14. Harlow DG (2014) Low cycle fatigue: probability and statistical modeling of fatigue life. In: Proceedings of the ASME 2014 pressure vessels and piping conference, PVP2014–28114, Anaheim, CA, 20–24 July 2014, V06BT06A045. ISBN 978–0–7918–4604–9
15. Harlow DG (2017) Confidence bounds for fatigue distribution functions. In: Proceedings of the ASME 2017 pressure vessels and piping conference, PVP2017–65416, Waikoloa, HI, 16–20 July 2017. V01AT01A011. ISBN 978–0–7918–5790–8
16. Thomas GB, Varma RK (1993) Evaluation of low cycle fatigue test data in the BCR/VAMAS intercomparison programme. Commission of the European Communities, BCR Information Applied Metrology, Brussels, Luxembourg
17. Harlow DG (2007) Probabilistic property prediction. Eng Fract Mech 74:2943–2951
18. Weibull EHW (1951) A statistical distribution function of wide applicability. ASME J Appl Mech 18:293–297
19. Abernethy RB (2006) In: Abernethy RB (ed) The new Weibull handbook, 5th edn. 536 Oyster Road, North Palm Beach, Florida, pp 33408–34328
20. Rinne H (2009) The Weibull distribution. CRC Press, Chapman-Hall, Taylor and Francis Group, Boca Raton

D. Gary Harlow is a Professor in Mechanical Engineering and Mechanics at Lehigh University in Bethlehem, PA. Since 2008 he has been the department chair. He received his BA in Mathematics (1973) from Western Kentucky University and his MS (1976) and Ph.D. in Applied Probability and Stochastic Processes (1977) from Cornell University. His research has primarily centered on scientifically and mechanistically based probability modeling and statistical analyses. Specifically, his research has included modeling of failure processes (creep, creep-fatigue, very high cycle fatigue, corrosion pitting, and corrosion fatigue) in materials (aluminum alloys, steels, and composites); stochastic fracture mechanics; mechanical and system reliability; applications of stochastic processes; and methods in applied statistics for materials characterization. His research has been sponsored throughout his career by the NSF, FAA, DoD, and DARPA. He is a fellow of ASME.

Reliability Improvement Analysis Using Fractional Failure

Mingxiao Jiang and Feng-Bin Sun

Abstract Many companies adopted Design for Reliability (DFR) process and tool sets to drive reliability improvement. However, it has been very challenging for business sector to conduct DFR with meaningful reliability metrics, when the reliability requirement is very high and the product samples in testing are very limited with small numbers of failure during product development tests. Bayesian approaches have been introduced by some companies to handle such challenges. Lately, some research has been conducted for reliability analysis by using fractional failures to count for failure fix effectiveness. In this paper, we will construct an approach to improve product reliability during development with fractional failure analysis method, incorporating failure fix effectiveness during each testing and failure fix phase. The fractional failure analysis is also expanded to accelerated reliability testing modeling.

Keywords Product development · Reliability improvement · Fractional failure · Fix effectiveness

1 Introduction

Due to increasing competition, decreasing development time and increasing product complexity, it is often difficult to meet customers' expectations of product performance, quality and reliability. To achieve high product quality ("out-of-box" product performance often quantified by Defective Parts Per Million), many companies successfully practice Design for Six Sigma (DFSS) to design high quality into products in early development stages. Similarly, to achieve high product reliability (often as measured by annualized failure rate, hazard rate, or mission reliability, etc.) some companies have introduced the DFR process to design high reliability into products in early development phase prior to transition into full production. DFR is intimately

M. Jiang (✉)
Medtronic, Plc., 7000 Central Ave NE, Minneapolis, MN 55432, USA
e-mail: mingxiao.jiang@medtronic.com

F.-B. Sun
Tesla, Inc., 3500 Deer Creek Road, Palo Alto, CA 94304, USA
e-mail: franksun9999@gmail.com

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_2

paced with various phases in the new product development cycle. Many companies serious about reliability are formally invoking DFR in the phase gate review process. DFR comprises the sequentially phased elements of reliability requirement setting, early prioritization of risk and allocation of reliability, design & optimization, validation, and control.

Even though a DFR process can clearly describe the steps and results needed to improve design (e.g., see Fig. 1 for a DFR process [1]), in practice though, it is very challenging to do so, since the reliability requirements could be very high, and the testing resources are limited. Bayesian reliability techniques have been adopted to deal with such challenges [1–3]. Lately, another kind of approach using fractional failure analysis (FFA) has been studied to incorporate fix effectiveness (FE) [4, 5] of corrective actions (CA) or countermeasures [6]. In the present work [7], FFA method is enhanced to count more complexities as in real product development process, including accelerated life testing (ALT) with competing failure modes (CFM) and combination of testing results in various design phases with improved product reliability.

This work starts with a brief introduction of a product development process, with reliability improvement needs. Next, the FFA, as a method to measure reliability, is briefly reviewed and expanded to deal with ALT analysis. Then, a new product

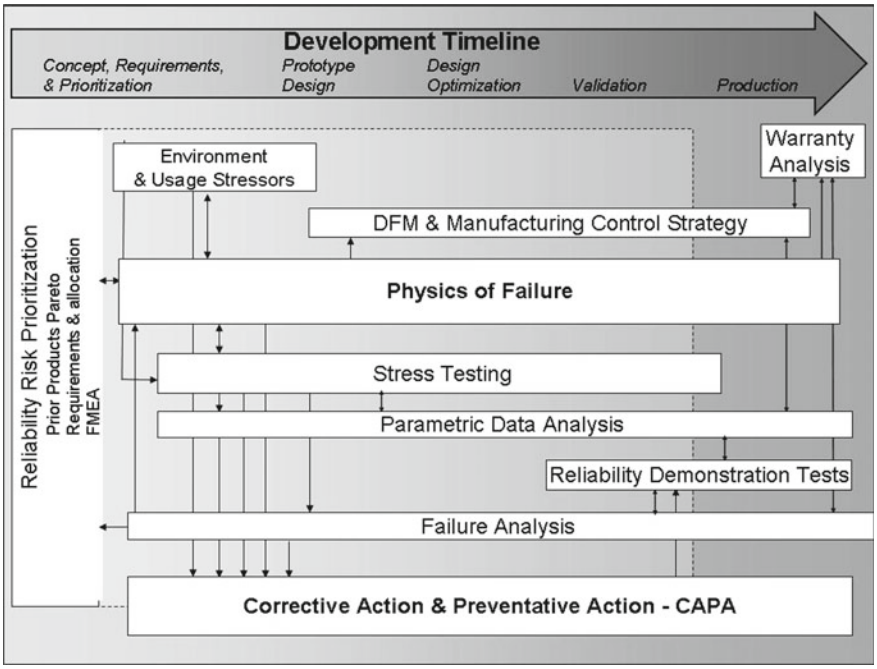


Fig. 1 DFR process during development

introduction (NPI) reliability improvement process with FFA is described, followed by an example to illustrate the proposed methodology.

2 Reliability Improvement with Fractional Failure Analysis

2.1 Product Development Process and Reliability Challenges

Product development starts from understanding customer and user needs. The whole development process can be illustrated by a V-diagram, e.g. the following one from many similar illustrations (Fig. 2):

During the process, the product is evolving from concept design, prototype design, improved and optimized designs, and verified and validated design. From reliability improvement perspective, the reliability testing and modeling can be conducted in each design phase (Fig. 3).

For example, Medical device manufacturers need to follow FDA Design Control regulations [8] during design and development process: user needs, design inputs, design process, design output, design verification, and design validation.

One of the most challenging tasks for reliability engineering during the product development process is to measure reliability improvement, especially when the test resource is very limited. In this situation, any early design and testing phase failures are very precious for failure mechanism understanding, failure mode correction/fix, and for reliability improvement modeling. In the next sections, an FFA approach is adopted to handle such a challenging reliability activity.

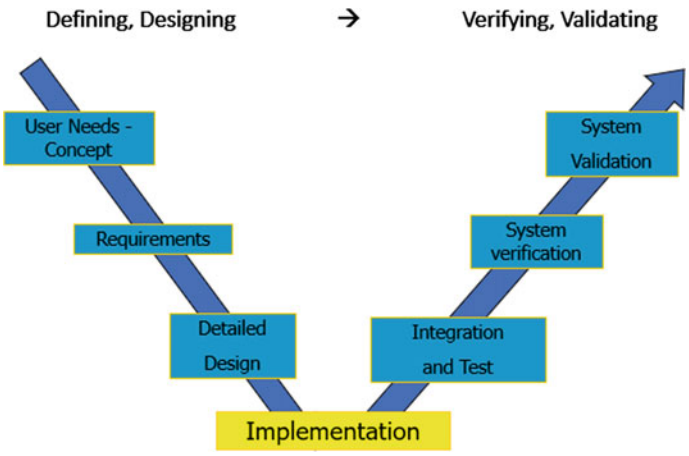


Fig. 2 Product development V-diagram

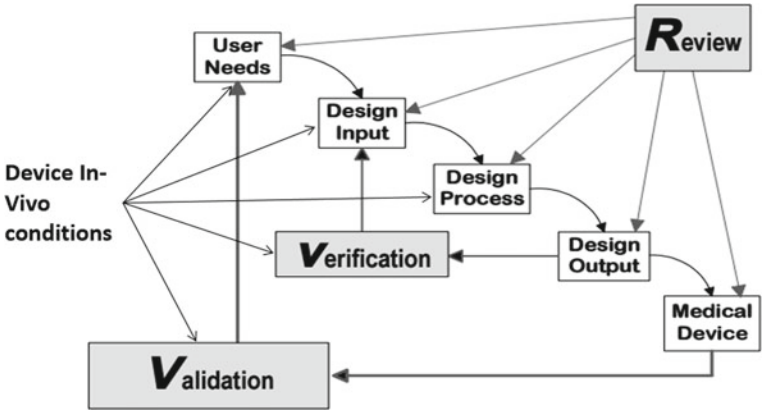


Fig. 3 Medical device development process [7]

3 Fractional Failure Analysis (FFA) and ALT

During product design and development, reliability engineers are often asked to project the reliability improvement based on reliability test (RT) data and potential effectiveness of corrective actions (CA) or counter measures (CM) on the identified failures. In practice, failures with CA in place are often discounted. However, not all CAs are effective, nor can they completely eliminate target failures. Additionally, their effectiveness is subject to verification. Reliability practitioners need to use caution in taking the CAs for granted. The corresponding data analysis and reliability assessment requires the introduction of fractional failures [4, 5].

Figure 4 illustrates the breakdown of a single failure considering the effectiveness of corrective actions. The likelihood (probability) of a failure with CA to be eliminated corresponds to the CA effectiveness, while the likelihood (probability) for this failure to recur corresponds to one minus CA effectiveness; i.e.,

$$\text{Failure Elimination Likelihood} = \text{CA Effectiveness} \tag{1}$$

$$\text{Failure Recurrence Likelihood} = 1 - \text{CA Effectiveness} \tag{2}$$

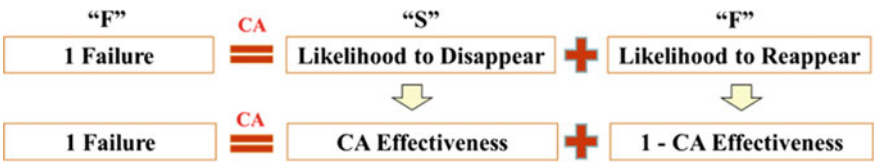


Fig. 4 Failure breakdown considering corrective action effectiveness

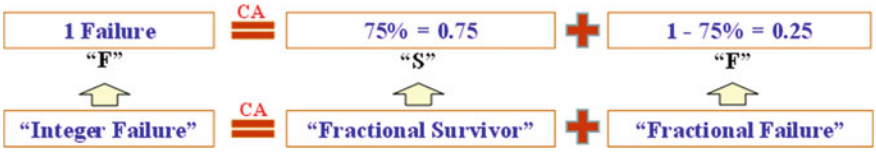


Fig. 5 An example of failure breakdown with 75% corrective action effectiveness

For example, if the overall effectiveness of all corrective actions developed for and applied to a particular failure root cause is 75%, then the probability for such failure to be eliminated is 0.75 while the probability of the failure to recur is 0.25, as shown in Fig. 5. After applying corrective actions, the previous failure thus becomes a fractional (0.75) survivor plus a fractional failure (0.25).

As shown in [5], the likelihood function of event observations with FFA is:

$$L(\theta_1, \ldots, \theta_k) = \left\{ \prod_{i=1}^{r} [f(T_i; \theta_1, \ldots, \theta_k)]^{\alpha_i} \prod_{i=1}^{r} [R(T_i; \theta_1, \ldots, \theta_k)]^{1-\alpha_i} \right\} \left\{ \prod_{i=1}^{m} [R(S_i; \theta_1, \ldots, \theta_k)]^{\delta_i} \prod_{i=1}^{m} [f(S_i; \theta_1, \ldots, \theta_k)]^{1-\delta_i} \right\} \left\{ \prod_{i=1}^{q} [F(I_{iU}; \theta_1, \ldots, \theta_k) - F(I_{iL}; \theta_1, \ldots, \theta_k)]^{\omega_i} \prod_{i=1}^{q} [R(I_{iU}; \theta_1, \ldots, \theta_k)]^{1-\omega_i} \right\} \quad (3)$$

where

$\theta_1, \theta_2, \ldots, \theta_k$	parameters of underlying life distribution
f	probability density function (pdf) of underlying life distribution
F	cumulative distribution function (CDF) of underlying life distribution
R	reliability function of underlying life distribution
k	total number of unknown distribution parameters
r	total number of confirmed whole or fractional failures at discrete time, T_i
q	total number of confirmed whole or fractional failures during time intervals
m	total number of confirmed whole or fractional failures at discrete time, S_i
T_t	effective operating time of ith failure, $i = 1, \ldots, r$
S_t	effective operating time of jth survivor, $i = 1, \ldots, m$
I_L	beginning time of interval
I_U	ending time of interval
α_i	failure confirmation rate of ith failure at T_i
$1 - \alpha_i$	proportion of ith failure that did not fail at T_i
δ_i	survivor confirmation rate of ith survivor at S_i
$1 - \delta_i$	proportion of ith survivor that did not survive at S_i

ω_i failure confirmation rate of ith failure in (I_{iL}, I_{iU})
$1-\omega_i$ proportion of ith failure that did not fail in (I_{iL}, I_{iU})

For ALT, the FFA likelihood function becomes,

$$L(\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c}) = \left\{\begin{array}{l}\prod_{i=1}^{r}[f(T_{i,j};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{\alpha_{i,j}} \\ \times\prod_{i=1}^{r}[R(T_{i,j};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{1-\alpha_{i,j}}\end{array}\right\} \left\{\begin{array}{l}\prod_{i=1}^{m}[R(S_{i,j};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{\delta_{i,j}} \\ \times\prod_{i=1}^{m}[f(S_{i,j};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{1-\delta_{i,j}}\end{array}\right\} \left\{\begin{array}{l}\prod_{i=1}^{q}[F(I_{(i,j)U};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c}) \\ -F(I_{(i,j)L};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{\omega_{i,j}} \\ \times\prod_{i=1}^{q}[R(I_{(i,j)U};\theta_1,\ldots,\theta_k,\theta_{k+1},\ldots,\theta_{k+c})]^{1-\omega_{i,j}}\end{array}\right\} \tag{4}$$

where, the second subscript j indicates the jth stress level, and $\theta_{k+1},\ldots,\theta_{k+c}$ are the ALT parameters to describe the stress life relationship.

4 Reliability Improvement Study Using FFA

During the product development process, multiple stages of reliability tests are conducted during different design phases from prototype design to final production. For each reliability test, all (major) failure modes are studied for root causes understanding and fixed through corrective actions. The fix effectiveness should be estimated/demonstrated for measuring reliability improvement. A simple documentation could be just as the following:

With information in Table 1, reliability assessment for each design phase can be quantified, using fractional failure/survivor approach:

- After RT2, conduct reliability assessment using both RT1 and RT2 data, with each failure in RT1 decomposed into a fractional failure and a fractional survivor, using Fig. 4 with "CA Effectiveness = FE_1".
- After RT3, conduct reliability assessment using RT1, RT2, and RT3 data:

Table 1 Fix effectiveness matrix

Failure modes	Fix	Fix Effectiveness for failures in RT1, FE_1	Fix Effectiveness for failures in RT2, FE_2

- Each failure in RT1 will be decomposed into a fractional failure and a fractional survivor, using Fig. 4 with "CA Effectiveness $= 1 - (1 - FE_1) * (1 - FE_2)$";
- Each failure in RT2 will be decomposed into a fractional failure and a fractional survivor, using Fig. 2 with "CA Effectiveness $= FE_2$".

5 Example

Reference [9] has competing failure mode (CFM) ALT data for a reliability test, which is referred as RT1 in this section. Assume the development team conducted failure analysis and implemented fixes to both failure modes. The FE is 10% for failure mode A and 90% for failure mode B. Then RT2 is conducted. After RT2, for both failure modes, additional understanding of the failure mechanism results in an estimate of FE as 50% for failure mode A and 80% for failure mode B during RT2. The data for RT2 and RT3 are from Monte Carlo simulation based on RT1 and fix effectiveness information. All the data are in Table 2 (in this section, all the time values are in hours):

Application of FFA [4, 5] to this reliability improvement process is conducted as follows, with analysis done by JMP 14:

- Step 1: CFM ALT analysis for RT1 (without fractional failures): This is the initial reliability level without fix, post RT1. Some results are given below, with reliability testing data under three different testing temperatures (300, 350 and 400 K) being fitted by an ALT model (Arrhenius stress-life relationship with Weibull distribution), and then projected to probability of failure as function of time under use condition (290 K). The results are plotted for the two different failure modes (A and B) separately in Fig. 6a, b.
- Step 2: CFM ALT analysis for the combination of RT1 (with fractional failures due to fix of the failures in RT1) and RT2 (without fractional failures): This is the reliability level post RT2. Some results are given below, again, with reliability testing data under three different testing temperatures being fitted by the same ALT model, and then projected to probability of failure as function of time under use condition, for both failure modes. Comparing Fig. 7 with Fig. 6, it can be seen that the sample size in Fig. 7 increases by combining RT1 and RT2 data, with RT1 data being treated as fractional failures and fractional survivals.

Table 2 ALT data

	RT1			RT2			RT3		
Test condition (Temp, K)	Failure mode	Time	Failure/Survival (F/S)	Failure mode	Time	F/S	Failure mode	Time	F/S
300	A	150	F	A	117	F	A	939	F
300	A	310	F	B	892	F	n/a	1000	S
300	A	393	F	A	216	F	n/a	1000	S
300	A	409	F	A	530	F	n/a	1000	S
300	B	421	F	A	870	F	n/a	1000	S
300	B	425	F	A	232	F	n/a	1000	S
300	B	459	F	A	910	F	n/a	1000	S
300	B	510	F	A	354	F	n/a	1000	S
300	A	512	F	A	818	F	n/a	1000	S
300	B	558	F	A	935	F	n/a	1000	S
300	B	568	F	A	948	F	n/a	1000	S
300	B	580	F	B	862	F	n/a	1000	S
300	B	615	F	B	557	F	A	766	S
300	A	616	F	A	39	F	n/a	1000	S
300	B	618	F	A	820	F	n/a	1000	S
300	B	629	F	A	863	F	n/a	1000	S
300	B	646	F	n/a	1000	S	n/a	1000	S
300	B	667	F	n/a	1000	S	n/a	1000	S
300	B	738	F	A	359	F	A	693	F
300	B	815	F	B	998	F	A	974	F

(continued)

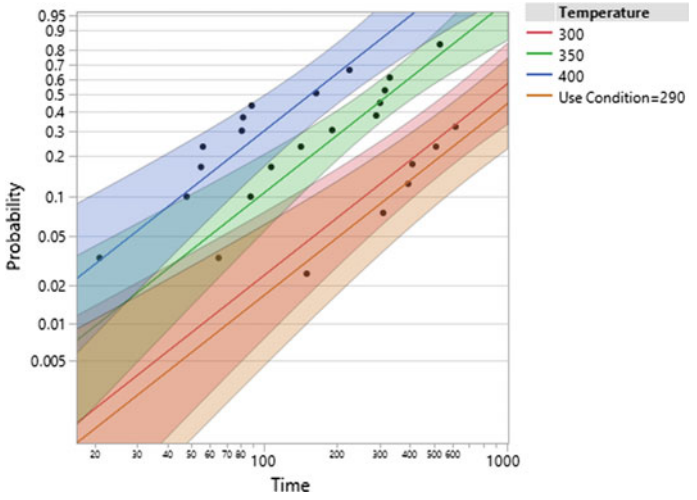
Table 2 (continued)

	RT1			RT2			RT3		
Test condition (Temp, K)	Failure mode	Time	Failure/Survival (F/S)	Failure mode	Time	F/S	Failure mode	Time	F/S
350	A	65	F	A	213	F	A	465	F
350	A	88	F	B	406	F	A	470	F
350	A	107	F	A	498	F	A	629	F
350	A	142	F	B	281	F	n/a	800	S
350	B	190	F	A	294	F	A	417	F
350	A	191	F	A	519	F	A	680	F
350	A	290	F	A	272	F	A	628	F
350	A	301	F	A	295	F	A	506	F
350	B	310	F	B	301	F	B	756	F
350	A	315	F	B	441	F	A	202	F
350	A	330	F	A	328	F	A	444	F
350	B	343	F	A	27	F	B	737	F
350	B	401	F	B	385	F	B	746	F
350	B	460	F	A	103	F	A	752	F
350	A	531	F	A	190	F	B	735	F
400	A	21	F	A	273	F	A	268	F
400	A	48	F	A	268	F	A	384	F
400	A	55	F	A	194	F	A	301	F
400	A	56	F	A	266	F	A	337	F

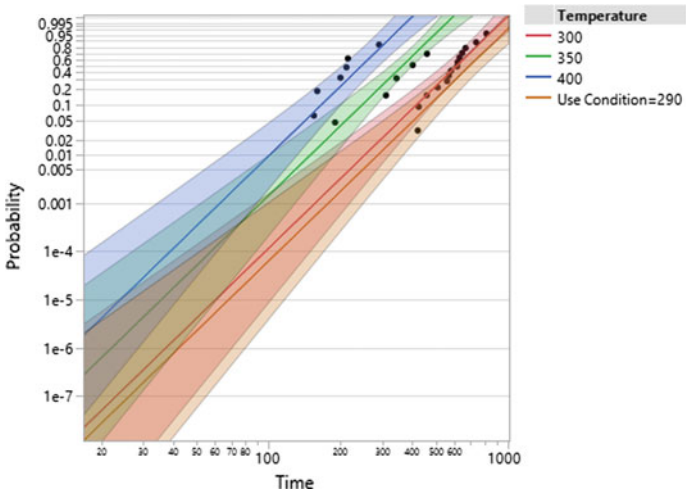
(continued)

Table 2 (continued)

	RT1			RT2			RT3		
Test condition (Temp, K)	Failure mode	Time	Failure/Survival (F/S)	Failure mode	Time	F/S	Failure mode	Time	F/S
400	A	81	F	B	400	F	A	305	F
400	A	82	F	A	119	F	B	134	F
400	A	89	F	A	323	F	A	483	F
400	B	155	F	A	173	F	A	369	F
400	B	160	F	A	120	F	B	250	F
400	A	164	F	A	39	F	A	186	F
400	B	200	F	A	82	F	A	272	F
400	B	212	F	A	165	F	A	314	F
400	B	215	F	A	260	F	B	345	F
400	A	225	F	B	289	F	A	411	F
400	B	290	F	A	392	F	B	500	F



(a) For FM A only

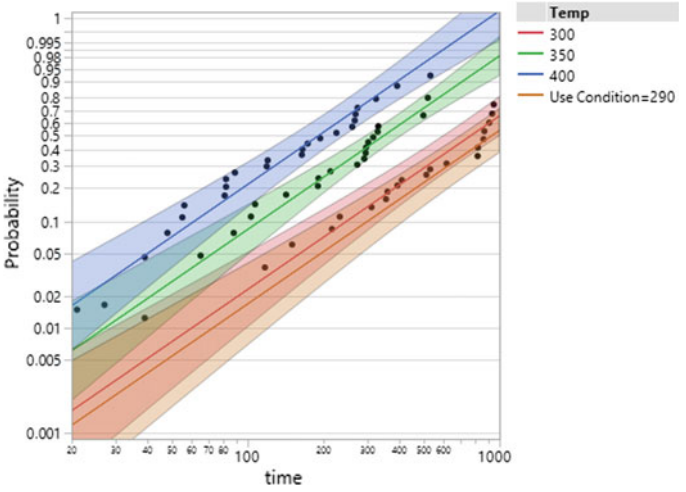


(b) For FM B only

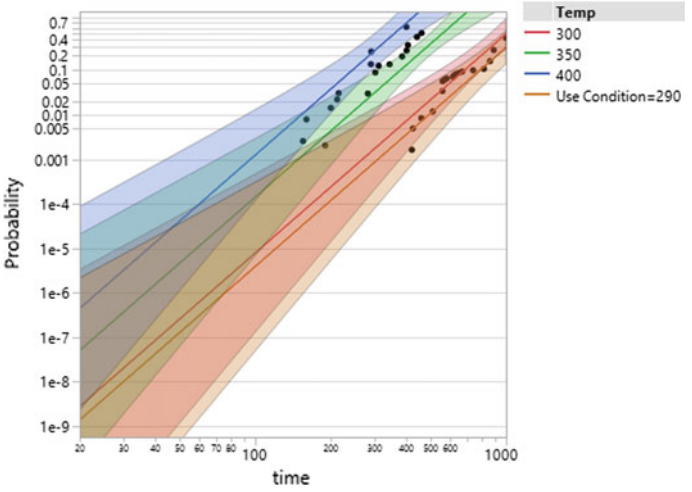
Fig. 6 ALT results for RT1

– Step 3: CFM ALT analysis for the combination of RT1, RT2 (both with fractional failures due to fix of failures) and RT3 (without fractional failures): This is the reliability level post RT3. Some results are given below. It can be seen that the sample size in Fig. 8 increases more by combining RT1, RT2 and RT3 data, with RT1 and RT2 data being treated as fractional failures and fractional survivals.

In each step above, the product reliability can be estimated by competing failure mode analysis:



(a) For FM A only



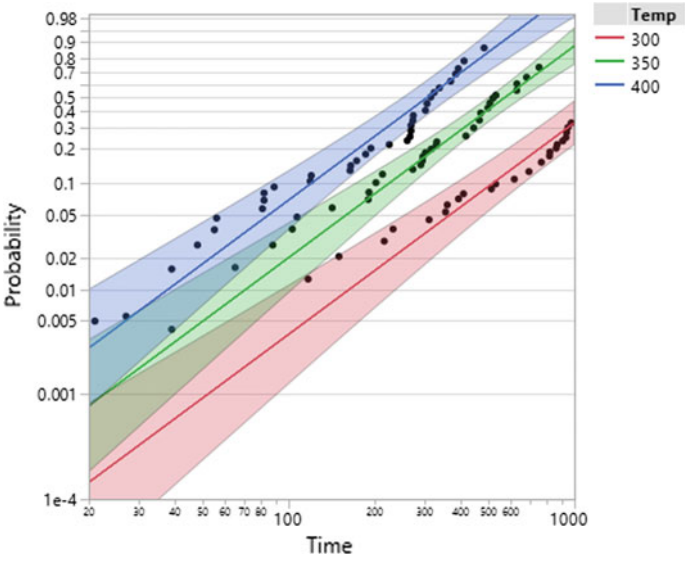
(b) For FM B only

Fig. 7 ALT results for RT1 (post fix) and RT2 combined

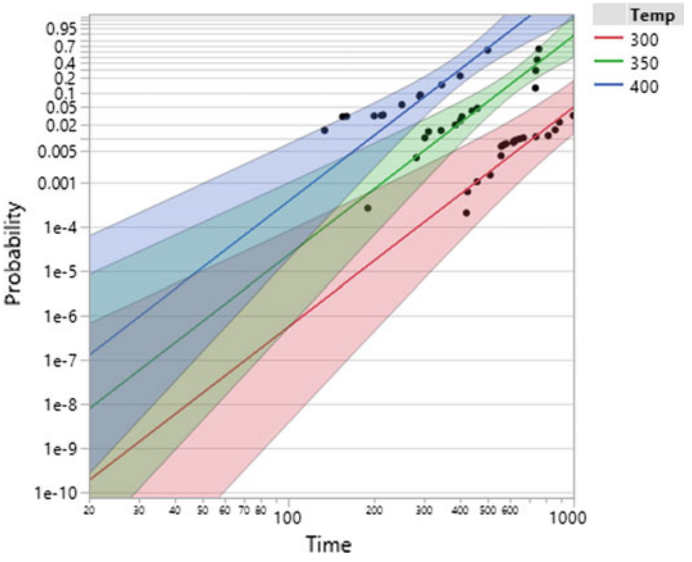
$$R(t|S) = R(t|S;\, A) \times R(t|S;\, B) \tag{5}$$

where, $t = 1200$ is product effective usage time, $S = 290$ K is use stress level, and A and B are failure modes. Results with such t and S are summarized in Table 3.

From Table 3, it can be observed that the overall reliability has been improving. From RT1 to post RT2, the reliability result against failure mode A does not increase, even though the true reliability against failure mode A should have increased. This is due to the fact that as competing failure mode B is getting fixed (fix of FM B is more



(a) For FM A only



(b) For FM B only

Fig. 8 ALT results for RT1

Table 3 Reliability improvement summary

FM A			
Design/Test phase	RT1	RT2	RT3
Activation energy	0.18	0.14	0.15
η	2235.1	1163.1	1936.8
β	1.54	1.66	1.82
FM B			
Design/Test phase	RT1	RT2	RT3
Activation energy	0.10	0.11	0.14
η	866.6	1243.0	2197.2
β	4.81	4.94	4.96
Design/Test phase	Predicted reliability for FM A	Predicted reliability for FM B	Predicted reliability for whole product
RT1	0.68	0.01	0.01
RT2	0.35	0.43	0.15
RT3	0.39	0.95	0.38

effective than fix of FM A in this example), failures of mode A (not fixed effectively) start to show up (used to be masked by failure mode B in RT 1).

As design and testing phase moves from RT1 to RT3, by counting the fractional failures/survivors in RT1 and RT2, the sample size increases in ALT analysis after RT2 from 50 (RT2 testing samples) to 100 (combined samples of RT1 (fractional failure/survivors) and RT2), and the sample size increases from 50 (RT3 testing samples) to 150 (combined samples of RT1 and RT2 (fractional failures/survivors) and RT3). The additive effect of fractional failures/survivors combining with RT3 samples can be seen in Fig. 9 (see the "Summary of Data" portion above "Scatterplot"), and the competing failure mode ALT analysis for failure mode B has results in Fig. 10.

The reliability confidence bounds are narrowing by counting fractional failures/survivors from RT1 and RT2, e.g., Fig. 11 is a comparison between RT1 analysis results (50 samples) and RT3 analysis results (with fractional failures and survivors from RT1 and RT2), focusing on FM B.

6 Discussion

In real product development, the system could be of much larger scale and more complex. Following the V-diagram, reliability target of the system can be allocated to subsystems, and with the reliability analysis approach illustrated previously, the reliability improvement can be tracked in Table 4.

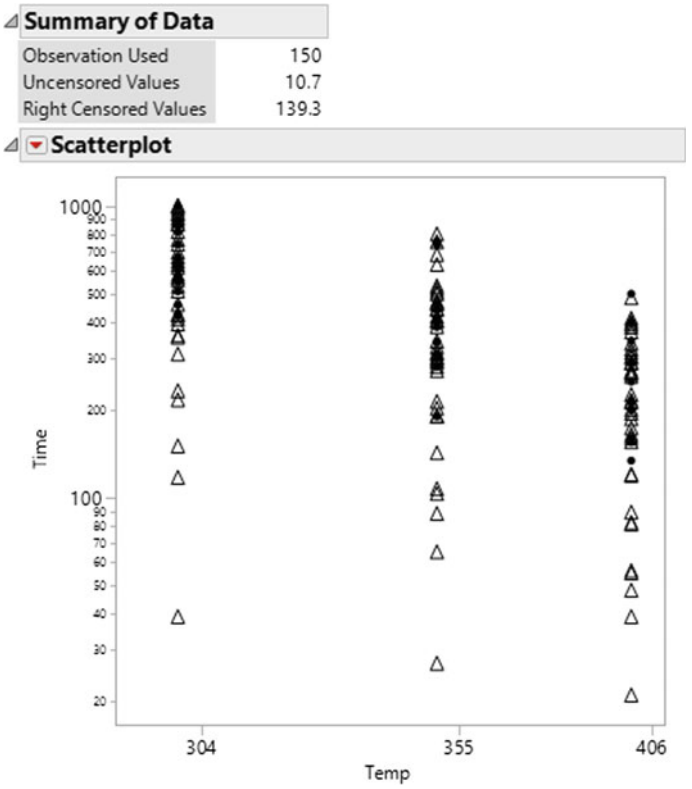


Fig. 9 ALT failures for mode B (with fractional failures/survivors from RT1 and RT2)

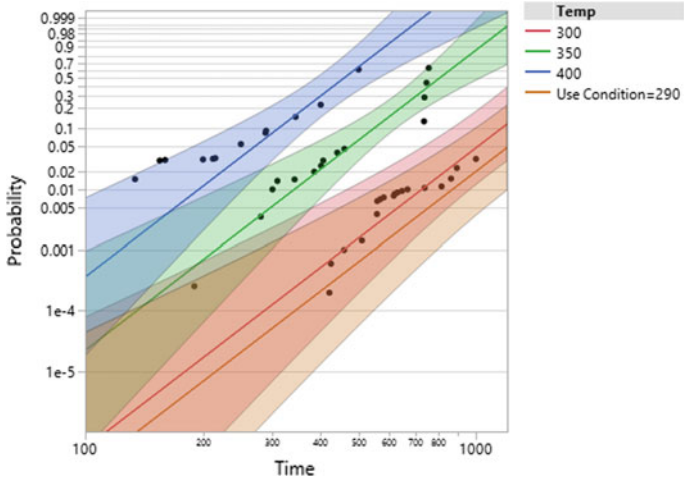


Fig. 10 ALT analysis for failure mode B with fractional failures/survivors from RT1 and RT2

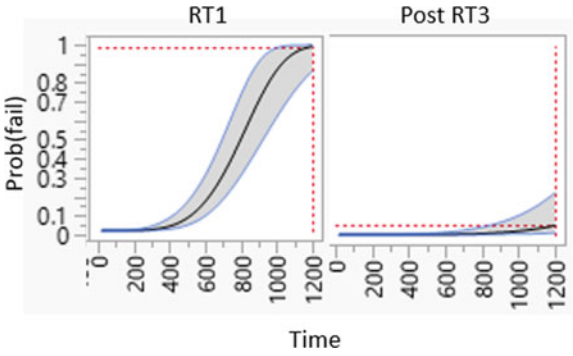


Fig. 11 Probability for failure mode B (comparing between RT1 and post RT3), under 290 K

Table 4 Documenting reliability improvement process

Subsystems	Allocated reliability targets	Major Failure Modes	Reliability for Major Failure Mode	Subsystem reliability (RT1)	Reliability improvement (fix post RT1)	FE post RT1	Subsystem reliability (RT2)	Reliability improvement (fix post RT2)	FE post RT2	Subsystem reliability (RT3)
Sub system 1										
Sub system 2										

The above reliability improvement can also be tracked using Bayesian reliability improvement method [10]. In such an approach, the reliability in a previous RT becomes the prior reliability for subsequent RT reliability data analysis.

References

1. Jiang M, Dummer D (2009) Enhancing Bayesian reliability demonstration test by design for reliability. Reliab Rev 29(4):18–24
2. Martz H, Waller R (1982) Bayesian reliability analysis. Krieger Publishing Company
3. Kleyner A et al (1997) Bayesian techniques to reduce the sample size in automotive electronics attribute testing. Microelectron Reliab 37(6):879–883
4. Sun F, Zhang J (2017) Dealing with fractional failures and survivors in data analysis of reliability engineering. In: Proceedings of 2017 annual reliability and maintainability symposium, Orlando, FL
5. Sun F, Jiang M (2018) Introduction to fractional failure and its application in reliability and risk management. In: Proceedings annual reliability and maintainability symposium, Reno, NV

6. Kurz D, Lewitschnig H, Pilz J (2014) Decision-theoretical model for failures which are tackled by countermeasures. IEEE Trans Reliab 63(2)
7. Jiang M, Sun F (2019) Reliability improvement with fractional failure analysis. In: Proceedings of the 25th ISSAT international conference on reliability and quality in design, Las Vegas, NV
8. Design control guidance for medical device manufacturers. An FDA Publication (1997)
9. Reliability Hot Wire (2006) 67
10. Jiang M (2018) Quantitative risk assessment with Bayesian reliability improvement. In: Proceedings of the 24th ISSAT international conference on reliability and quality in design, Toronto, Canada

Mingxiao Jiang received his B.S. in Engineering Mechanics from Zhejiang University, M.S. in Reliability Engineering from University of Arizona, and Ph.D. in Mechanics of Materials from Georgia Institute of Technology. He is currently a Technical Fellow and a Senior Principal Reliability Engineer at Medtronic. He is a Certified Reliability Engineer, and senior members of ASQ and IEEE. He has 3 US patents and 50 publications in referred journals and conference proceedings. He has been active in reliability societies as a Vice Chair for organization RAMS® (2009 to 2013) and committee members for several international reliability conferences. He has been serving on the Editorial Review Boards for Reliability Review and Quality Engineering.

Feng-Bin (Frank) Sun is currently a senior staff of Reliability Engineering, technical lead of manufacturing reliability, and manager of Design for Reliability (DFR) at Tesla Inc. with over 30 years of industry and academia experience. Prior to joining Tesla, he worked at HGST, a Western Digital Company, as a Senior Technologist of Reliability Engineering from 2010 to 2015, at Western Digital Corporation as a Technical Team Leader from 2003 to 2010, and at Maxtor-Quantum Corporation as a senior staff Reliability Engineer from 1997 to 2003. He has published two books, both by Prentice Hall, and more than 40 papers in various areas of reliability, maintainability and quality engineering. He received his B.S. degree in Mechanical Engineering from the Southeast University (China), M.S. degree in Reliability from the Shanghai University (China), and Ph.D. in Reliability from the University of Arizona. Dr. Sun served in the editorial board as an Associate Editor for the IEEE Transactions on Reliability from 1999 to 2003 and from 2016 to present, Program Chair of ISSAT International Conference on Reliability and Quality in Design since 2014, and a committee member and session moderator of numerous international conference on reliability. He is a senior member of ASQ and the President for Society of Reliability Engineers (SRE) Silicon Valley Chapter. He is also the program manager in charge of ASQ Reliability Webinar Mandarin Series. Dr. Sun received the RAMS 2013 A.O. Plait Best Tutorial Award.

Modelling Innovation Paths of European Firms Using Fuzzy Balanced Scorecard

Petr Hájek, Jan Stejskal, Michaela Kotková Stříteská, and Viktor Prokop

Abstract Because innovation processes are complex, uncertain and highly dimensional, modelling innovation paths is a very challenging task. As traditional regression models fail to address these issues, here we propose a novel approach for the modelling. The approach integrates Balanced Scorecard, a method used for strategic performance measurement, and fuzzy set qualitative comparative analysis. In addition to key performance indicators, strategic goals are taken into consideration in the modelling. We provide empirical evidence for the effectiveness of the approach on a large dataset of European firms. We show that several innovation pathways can be identified for these firms, depending on their strategic goals. These results may be of relevance for the decision making of innovative firms and other actors of innovation system.

Keywords Innovation · Performance measurement · Balanced scorecard · Fuzzy sets · Qualitative comparative analysis

1 Introduction

Business managers have been trying to increase the competitive advantage of both businesses and entire industrial sectors over the last 100 years. It was proved (by the developments of scientific knowledge, and also by the practice), that the firm's competitive advantage depends not only on the availability and the volume of capital. Nowadays, the competitiveness of a firm depends also on a number of other "soft"

P. Hájek · J. Stejskal (✉) · M. Kotková Stříteská · V. Prokop
Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic
e-mail: jan.stejskal@upce.cz

P. Hájek
e-mail: petr.hajek@upce.cz

M. Kotková Stříteská
e-mail: michaela.striteska@upce.cz

V. Prokop
e-mail: viktor.prokop@upce.cz

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_3

inputs such as knowledge, quality, creativity, learning ability, ability to cooperate and use different factors in the minimum time and cost savings. These factors are crucial to the ability of an enterprise to create innovative products and succeed in international (but often in national) markets. Therefore, the performance (or competitiveness) management of innovation activities is the topic of many studies [13]. Specifically, researchers and managers are interested in which of the factor in socio-economic environment has a significant positive effect on the innovation performance.

Scholars agree on the importance of organizational innovation for competitiveness. This form of innovation is perceived as prerequisite as well as facilitators and mainly also as immediate sources of competitive advantage in the business environment [8]. Hult et al. [18] stated that organizations' culture competitiveness is based on four factors—entrepreneurship, innovativeness, market orientation and organizational learning. Study of [29] confirmed that organizational culture, the environment embedded in corporate strategy has a positive impact on performance. Company strategy should be the result of strategic planning and application of various aggregated activities. Value chain is created, in whose conceptual structure it is possible to analyze the sources of competitive advantage [32].

Thanks to the rapid ability of the firms to introduce innovation and to high pressure to reduce costs, firms must increasingly assess their performance and be subject to benchmarking [9]. There is a growing need for continuous improvement. This principle must be an integral part of strategic management activities and it is also one of the parameters of multifaceted performance [33]. Managers' efforts to balance corporate strategies, the need for detailed knowledge of production inputs, and the precise knowledge of customer requirements have resulted in the use of the Balanced Scorecard (BSC) model. This familiar tool is helping to balance the firm strategy and define performance multifaceted indicators.

When we want to measure a multifaceted performance we need to use non-financial indicators which are not free of limitations. As [6] stated, some non-financial performance indicators may be difficult to measure accurately, efficiently, or in a timely fashion. According to Ittnerr and Larcker [19], other limitations are that they may be biased or ambiguous, easier to manipulate, measured in many ways that may change over time, time-consuming, and expensive. Because innovations most of the time track intangible assets, their measurement remains a challenge for most of the enterprises as well as academics [20]. To date, no rigorous model for innovation performance measurement has been found due to complexity and uncertainty present in innovation processes [14]. To overcome these problems, here we develop an integrated approach that combines strategic innovation performance assessment with qualitative comparative analysis. This approach enables finding innovation pathways within the BSC framework, thus connecting the performance criteria with corporate strategic priorities.

The remainder of this paper is structured as follows. In the next section, we present the theoretical background for BSC model and its use in the strategic management of innovation processes in companies. Section 3 provides the characteristics of the dataset and the research methodology. Section 4 presents the experimental results.

In Sect. 5, we discuss the results that were obtained and conclude the paper with suggestions for future research.

2 Theoretical Background

In a rapidly changing environment, enterprises must continuously innovate, develop new processes and deliver novel products or services to achieve and maintain a competitive advantage [36]. Hence nowadays, it is not about whether to innovate or not, but how to innovate successfully. In this context, innovation management itself is evolving and presents enterprises with tough challenges in performance measurement [7].

According to [10] most enterprises do not measure the benefits created by their innovation projects, many of them do not have internal structures to measure innovation and do not pay sufficient attention to the process of innovation management. This fact is confirmed by a number of surveys, for example The Boston Consulting Group's survey revealed that 74% of the executives believed that their company should track innovation as rigorously as core business operations, but only 43% of the companies actually measured innovation [2]. Dewangan and Godse [7] state that less than 41% of enterprises regard their Innovation Performance Measurement (IPM) systems as effective and a large majority of enterprises have felt the need to improve their IPM system. According to [21] the real issue is a lack of metrics and measurements which cause that companies measure too little, measure the wrong things or not measure innovation at all.

The main reason why a number of enterprises are still struggling with problems in this area is that measurement of innovation is intangible by nature. Innovations typically create much more intangible than tangible value, and intangible value cannot be measured using traditional financial methods [10]. Adams et al. [1] stress the absence of the frameworks for innovation management, measurement indicators as well as the relatively small number of empirical studies on measurement in practice. Still, we can find studies dealing with innovation performance measurement [15].

The measurement of innovation is also crucial issue for an academic research. As stated Adams et al. [1] unless constructs relating to the phenomenon are measurable using commonly accepted methods, there is a risk that different operationalization of the same effect will produce conflicting conclusions, and that theoretical advances become lost in the different terminologies.

In literature we can find three types of innovation performance models that are based on different methodologies including empirical ones like firms survey [30], case study [25] and theoretical approaches [38]. Recent frameworks used for innovation measurement are also diverse and often built with operations research tools such as Data Envelopment Analysis (DEA) or multi-criteria analysis tools such as Analytic Hierarchy Process (AHP) [28]. DEA is appropriate for innovation efficiency calculation and for benchmark [34, 35] and AHP for innovation portfolio management [24]. However, traditional models and frameworks for innovation measurement

are mainly oriented on past performance and stressing more control than process of innovation management.

Currently Balanced Scorecard is considered the most comprehensive model for the continuous improvement of the organization's performance. BSC can be used also to measure the performance of the innovation processes. The authors of the BSC concept itself state that the recent BSC provides a richer, more holistic view of the organization, and the model can be used with any selection of perspective appropriate to a particular exercise [17]. The fact that it can be adapted according to the needs of the organization in any of its areas, makes from BSC the most appropriate tool for introducing a complex system of measuring innovation performance [10, 15, 20].

In the context of innovation measurement a key improvement of BSC is the linkage of measurement to a strategy map; this tighter connection between the measurement system and the strategy map elevates the role of nonfinancial measures in strategy implementation and strategy evaluation [22]. Strategic alignment determines the value of intangible assets and strategy map is designed to help execute strategy and bring predictive qualities to key performance indicators [5]. Only by considering the cause-effect relationships, it is possible to switch from innovation performance measurement to innovation performance management.

One of the first authors that suggested BSC as an integrated performance measurement system for research and development were Kerssens-van Dronglene and Cook [23]. Later Bremser and Barsky [4] extended their work by integrating Stage-Gate approach to R&D management with the BSC to present a framework to show how firms can link resource commitments to these activities and strategic objectives. Gama et al. [10] proposed an Innovation Scorecard based on innovation metrics and the traditional BSC in order to measure the value added by innovation and tested the pilot in a large industrial company. Saunila and Ukko [37] introduce a conceptual framework of innovation capability measurement based on the literature review of performance measurement frameworks and assessment models. Ivanov and Avasilcăi [20] developed new model on the basis of detailed analysis of four most important performance measurement models (BSC, EFQM, Performance Prism and Malcolm Baldrige) that tries to emphasize the most important characteristics that have to be analyzed when innovation performance is measured. Zizlavsky [44] proposed an Innovation Scorecard based specifically on project management, BSC, input-process-output-outcomes model and Stage Gate approach. Above presented models and frameworks are theoretical or based on the analysis of one or few case studies and the results have not been verified on a large data set.

When using non-financial measures, we need to often deal with the data vagueness and ambiguity. The contemporary research reveals that the fuzzy set and the fuzzy logic theory are the appropriate theoretical background to solve this issue. Recently, many researchers have been developed and modified fuzzy linguistic approach in order to apply in diverse domains, some of them integrated fuzzy linguistic also in BSC structure [15, 27].

In one of the oldest works are BSC and Fuzzy logic combined into a methodology and a software tool that is able to help executives to optimize the strategic business processes [12]. Later on, several scholars integrates BSC with Fuzzy AHP to calculate

the relative weights for each performance measure [26, 39]. More recent studies propose approaches to analyze strategy map using BSC-FAHP to asses each aspects of strategy or impact of changes in the mission and vision of the organization [31, 40]. Interesting is also the approach of [3] that provide a first semantic fuzzy expert system for the BSC. Its knowledge base relies on an ontology and its inference system derives new knowledge from fuzzy rules.

However, none of them, to the best of our knowledge, utilized the advantages of fuzzy sets to model innovation paths within the BSC framework.

3 Research Methodology

In this section, we first introduce a BSC for evaluating firm innovation performance. Specifically, we adapted the model proposed in [13] that provides a set of indicators for each perspective of BSC. Our contribution is the inclusion of vision and strategy effect on all four BSC perspectives (Fig. 1). More precisely, five strategies were considered: (1) developing new markets; (2) reducing costs; (3) introducing or improving goods or services and their marketing; (4) increasing organization flexibility; and (5) building alliances with other firms and institutions. The causal effects in the BSC model also represent the theoretical framework for further detection of innovation paths.

The indicators used to evaluate firm innovation performance and their innovation strategies are presented in Table 1. Note that both quantitative and qualitative indicators were included and that these were fuzzified to [0, 1] interval for further qualitative comparative analysis. Note that the values are regularly distributed in the [0, 1] to make the fuzzification as objective as possible. The source of the data was

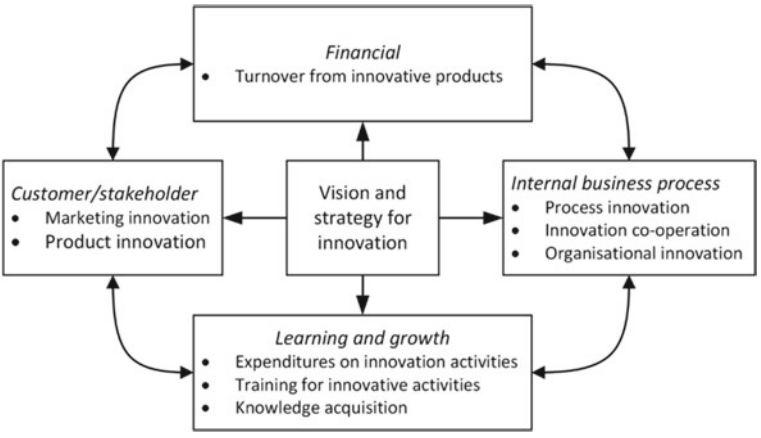


Fig. 1 BSC for evaluating firm innovation performance

Table 1 Indicators for BSC model

	Indicator	Description
l_1	Training for innovative activities	0 for no, 0.5 for medium and 1 for high intensity
l_2	Expenditure on innovation activities	Rescaled to [0, 1]
l_3	Knowledge acquisition	Importance of knowledge sources (0 for no importance, 0.25 for low, 0.5 for medium and 1 for high importance)
c_1	Product innovation	0 for none, 0.5 for product new to the firm and 1 for product newt to the market
c_2	Marketing innovation	0 for no marketing innovation, 0.25 for one type of marketing innovation, 0.5 for two of them, 0.75 for three of them, and 1 for the implementation of all types of marketing innovations
i_1	Process innovation	0 for no process innovation, 0.33 for one type of process innovation, 0.67 for two of them, and 1 for the implementation of all types of process innovations
i_2	Innovation co-operation	Diversity of cooperation partners: 0.14 for 1 partner, 0.29 for 2 partners, 0.43 for 3, 0.57 for 4, 0.71 for 5, 0.86 for 6, and 1 for 7 cooperation partners
i_3	Organizational innovation	0 for no organizational innovation, 0.33 for one type of organizational innovation, 0.67 for two of them, and 1 for the implementation of all types of organizational innovations
f_1	Turnover from innovative products	The share of turnover scaled to [0, 1]
s_1	Developing new markets	Importance of strategies: 0 for not relevant, 0.33 for low importance, 0.67 for medium importance and 1 for high importance
s_2	Reducing costs	
s_3	Introducing (improving) goods or services	
s_4	Increasing flexibility	
s_5	Building alliances	

the CIS (Community Innovation Survey) harmonized questionnaire, for details see the questionnaire available at: http://ec.europa.eu/eurostat/documents/.

The data used in this study come from the CIS survey for the period 2010–2012. To collect the data, CIS is based on the combination of exhaustive surveys and stratified random sampling of firms with at least ten employees. Overall, we were able to collect the data for 17,586 firms. The following countries were included in the dataset: Germany (5,777 firms), Portugal (3,341), Bulgaria (2,409), Hungary (1,182), Croatia (944), Romania (829), Estonia (771), Slovenia (733), Lithuania (653), Slovakia (560) and Cyprus (388). To impute missing data, median values of the respective industry was applied [28, 43].

As the data available at the Eurostat upon request, we report only their main characteristics here. Most firms in the dataset were SMEs (small and medium enterprises), introducing mostly product innovations out of the four innovation types. About half of the firms were innovative, with more than 20% turnover from innovative products. Suppliers were the most frequent cooperation partners. Regarding the innovation strategies, reducing costs was the most important one and increasing flexibility placed second. In contrast, building alliances and developing new markets were the least important strategies.

To model the innovation paths, we adopted the approach used in earlier studies for this task [11, 41] and used fuzzy set qualitative comparative analysis [42]. This method is suitable for modelling complex decision-making processes and detects also asymmetrical dependencies in the data [42]. This is done by examining all possible input-output combinations. Thus, necessary and sufficient conditions can be identified for each output. Those paths that are consistent with the data (with conditions covered by the output) and that cover sufficient data instances are stored in the solution. To calculate the fuzzy set membership of the path configuration, Gödel's *t*-norm was used for logical AND, this is the minimum operation was performed over the conditions. To obtain the consistency (cons) and coverage (cov) of the solutions, the following equations were used [42]:

$$cons(x_i \leq y_i) = \sum_{i=1}^{n} \min(x_i, y_i) / \sum_{i=1}^{n} x_i, \tag{1}$$

$$cov(x_i \leq y_i) = \sum_{i=1}^{n} \min(x_i, y_i) / \sum_{i=1}^{n} y_i. \tag{2}$$

4 Experimental Results

To perform the modelling of innovation paths, we used the freely available fsQCA software. For the sake of space, here we present only the most important path configuration, with the financial perspective f_1 as the output and the remaining perspectives and strategies as the input variables. Consistency cutoff was set to 0.5 to discard less consistent pathways.

The results of the modelling included 11 paths as follows:

Path 1: c_1 and ~c_2 and ~l_1 and ~l_3 and ~i_1 and ~i_2 and ~i_3 and s_1 and s_2 and s_3 and s_4 with cov = 0.20 and cons = 0.51,
Path 2: c_1 and ~c_2 and ~l_1 and ~i_1 and ~i_2 and ~i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.15 and cons = 0.52,
Path 3: c_1 and c_2 and l_1 and l_3 and i_1 and i_2 and i_3 and s_2 and s_3 and s_4 with cov = 0.10 and cons = 0.57,

Path 4: c_1 and c_2 and l_1 and l_3 and i_1 and i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.10 and cons = 0.57,
Path 5: c_1 and ~c_2 and ~l_1 and ~l_2 and ~l_3 and ~i_1 and ~i_2 and ~i_3 and s_1 and s_2 and s_3 and ~s_5 with cov = 0.11 and cons = 0.50,
Path 6: ~c_2 and ~l_1 and ~l_2 and l_3 and ~i_1 and i_2 and ~i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.09 and cons = 0.55,
Path 7: c_1 and c_2 and l_1 and l_2 and l_3 and i_1 and i_2 and i_3 and s_1 and s_2 and s_3 and s_4 with cov = 0.08 and cons = 0.64,
Path 8: c_1 and l_1 and l_2 and l_3 and i_1 and i_2 and i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.09 and cons = 0.63,
Path 9: c_1 and c_2 and l_1 and l_2 and l_3 and i_2 and i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.08 and cons = 0.64,
Path 10: c_1 and ~c_2 and ~l_1 and l_2 and ~l_3 and ~i_1 and ~i_2 and ~i_3 and ~s_1 and ~s_2 and ~s_3 and ~s_4 and ~s_5 with cov = 0.04 and cons = 0.57,
Path 11: c_1 and ~c_2 and l_1 and l_2 and l_3 and ~i_1 and i_2 and ~i_3 and s_1 and s_2 and s_3 and s_4 and s_5 with cov = 0.06 and cons = 0.62,

where cov denotes coverage, cons is consistency and ~ indicates negation operator $\sim x = 1 - x$. For the complete solution of the model, we obtained the consistency of 0.52, which can considered sufficient relative to previous literature [26]. In addition, the solution had a high coverage of 0.39.

Path 1 and Path 5 state that high turnover from innovative products can be achieved by introducing product innovation only. However, those paths are not suitable for firms aiming to build strategic alliances. Overall, we can state that product innovation is a necessary condition to yield financial profit from innovation activities. As suggested by Path 2, this holds irrespective of the strategic goals. What is more interesting is the fact in Path 6 this condition is not met, and product innovation is replaced by knowledge acquisition and innovation co-operation by those firms. Firms following Path 3 introduced all the four types of innovations during the monitored period, which required both the innovation co-operation and training for innovative activities combined with knowledge acquisition. Path 4 represents a similar solution but for firms seeking for collaborative partners and new markets.

Interestingly, expenditure on innovation activities was not a necessary condition to achieve high share of turnover from innovative products in Paths 1–6, suggesting that these represent the configuration paths taken by SMEs. In contrast, the expenditure was present in Paths 7–11, which provided more consistent solutions. We can observe that high expenditure on a firm's innovation activities is often accompanied with intensive training and knowledge acquisition, this is a coordinated effort for learning and future growth. Also note that Path 10 suggests that the five strategies included in the data did not provide an exhaustive listing. Additional strategies can be considered, such as reducing consumption of raw materials or improving working conditions.

To further study the firms taking the identified path configurations, we investigated their countries of origin and industry classification. For the former, we classified the countries according to the EIS (European Innovation Scoreboard) categories into modest innovators, moderate innovators, strong innovators and innovation leaders.

Regarding the latter, we used the Eurostat definition of knowledge-intensive industries (i.e., those for which tertiary educated employees represent more than 33% of the industry employment). Based on these classifications, we found that Paths 1, 5 and 6 were taken by firms from less performing countries and knowledge-non-intensive industries. Paths 3, 4, 7 and 10 represent knowledge intensive industries from less performing countries. Finally, Paths 8, 9 and 11 were taken by firms from well performing countries and knowledge-intensive industries primarily.

5 Conclusion

In conclusion, we proposed a novel approach to modelling innovation paths that integrates innovation BSC with qualitative comparative analysis. We have outlined the BSC model, taking into account both the importance of firms' innovation strategies and the qualitative assessment of many innovation indicators. A large dataset of European firms was used to validate our approach. The findings of this study imply that strategic priorities are crucial for the selection of firms' innovation pathways. In addition, the results suggest that are eleven consistent innovation paths present in the dataset.

Prior research has indicated that expenditure on innovation activities is the most important determinants of innovation outcomes [16]. Indeed, we demonstrated that the expenditure is a necessary condition in highly innovative countries. However, the economic outcomes of innovation activity can also be achieved by introducing product innovation in less performing countries, when appropriate innovation strategy is taken. We believe that this research will serve as a base for future studies focused on individual countries and industries. The model proposed here may be applied to different CIS datasets without limitations. Moreover, the set of the BSC indicators, as well as the fuzzification procedure, may be adjusted to consider country/industry specifics. On the one hand, the variety of firms included in our dataset provided an extensive empirical support to our approach. On the other hand, more consistent results can be achieved in future studies when investigating a more specific sample of firms.

Acknowledgements This work was supported by a grant provided by the scientific research project of the Czech Sciences Foundation Grant No. 17-11795S.

References

1. Adams R, Bessant J, Phelps R (2006) Innovation management measurement: a review. Int J Manag Rev 8(1):21–47
2. Andrew JP, Manget J, Michael DC, Taylor A, Zablit H (2010) Innovation 2010: a return to prominence—and the emergence of a new world order. The Boston Consulting Group, pp 1–29

3. Babillo F, Delgado M, Gómez-Romero J, López E (2009) A semantic fuzzy expert systome for a fuzzy balanced scorecard. Expert Syst Appl 36:423–433
4. Bremser WG, Barsky NP (2004) Utilizing the balanced scorecard for R&D performance measurement. R&D Manag 34(3):229–238
5. Buytendijk F, Hatch T, Micheli P (2010) Scenario-based strategy maps. Bus Horiz 53:335–347
6. Chow ChV, Van der Stede WA (2006) The use and usefulness of nonfinancial performance measures. Manag Account Q 7(3):1–8
7. Dewangan V, Godse M (2014) Towards a holistic enterprise innovation performance measurement system. Technovation 34(9):536–545
8. Donate MJ, de Pablo JDS (2015) The Role of knowledge-oriented leadership in knowledge management practices and innovation. J Bus Res 68(2):360–370
9. Estampe D, Lamouri S, Paris JL, Brahim-Djelloul S (2015) A framework for analysing supply chain performance evaluation models. Int J Prod Econ 142(2):247–258
10. Gama N, Silva MM, Ataíde J (2007) Innovation scorecard: a balanced scorecard for measuring the value added by innovation. In: Cunha PF, Maropoulos PG (eds) Digital enterprise technology. Springer, Boston, MA, pp 417–424
11. Ganter A, Hecker A (2014) Configurational paths to organizational innovation: qualitative comparative analyses of antecedents and contingencies. J Bus Res 67(6):1285–1292
12. Haase VH (2000) Computer models for strategic business process optimisation. In: Proceedings of the 26th euromicro konference, vol 2, pp 254–260
13. Hajek P, Henriques R (2017) Modelling innovation performance of european regions using multi-output neural networks. PLoS ONE 12(10):e0185755
14. Hajek P, Henriques R, Castelli M, Vanneschi L (2019) Forecasting performance of regional innovation systems using semantic-based genetic programming with local search optimizer. Comput Oper Res. https://doi.org/10.1016/j.cor.2018.02.001 (2019)
15. Hajek P, Striteska M, Prokop V (2018) Integrating balanced scorecard and fuzzy TOPSIS for innovation performance evaluation. In: PACIS 2018 proceedings, pp 1–13
16. Hashi I, Stojčić N (2013) The impact of innovation activities on firm performance using a multi-stage model: evidence from the community innovation survey 4. Res Policy 42(2):353–366
17. Hoque Z (2014) 20 years of studies on the balanced scorecard: trends, accomplishments, gaps and opportunities for future research. Br Account Rev 46:33–59
18. Hult GTM, Snow CC, Kandemir D (2003) The role of entrepreneurship in building cultural competitiveness in different organizational types. J Manag 29(3):401–426
19. Ittner CD, Larcker DF (2011) Assessing empirical research in managerial accounting: a value-based management perspective. J Account Econ 32:349–410
20. Ivanov CI, Avasilcăi S (2014) Measuring the performance of innovation processes: a balanced scorecard perspective. Procedia Soc Behav Sci 109:1190–1193
21. James PA, Knut H, David CM, Harold LS, Andrew T (2008) A BCG senior management survey—measuring innovation 2008: squandered opportunities (2008)
22. Kaplan RS, Norton DP (2014) The strategy map: guide to aligning intangible assets. Strat Leadership 32(5):10–17
23. Kerssens-van Drongelen IC, Cook A (2007) Design principles for the development of measurement systems for research and development process. R&D Manag 27(4):345–357
24. Khaksari S (2017) AHP and innovation strategy as project portfolio management. Polytechnic University of Turin (2017)
25. Lazzarotti V, Manzini R, Mari L (2011) A model for R&D performance measurement. Int J Prod Econ 134(1):212–223
26. Lee AHI, Chen W-CH, Chang Ch-J (2008) A fuzzy AHP and BSC approach for evaluating performance of IT department in the manufacturing industry in Taiwan. Expert Syst Appl 34:96–107 (2008)
27. Lin QL, Liu L, Liu H-C, Wang DJ (2013) Integrating hierarchical balanced scorecard with fuzzy linguistic for evaluating operating room performance in hospitals. Expert Syst Appl 40:1917–1924

28. Mbassegue P, Nogning FL, Gardoni M (2016) A conceptual model to assess KM and innovation projects: a need for an unified framework. In: Bouras A, Eynard B, Foufou S, Thoben KD (eds) Product lifecycle management in the era of internet of things. PLM 2015. IFIP advances in information and communication technology. Springer, Cham, vol 467
29. Naranjo-Valencia JC, Jiménez-Jiménez D, Sanz-Valle R (2016) Studying the links between organizational culture, innovation, and performance in Spanish companies. Revista Latinoamericana de Psicología 48(1):30–41
30. OECD (2005) OECD SME and entrepreneurship outlook 2005. OECD Publishing, 416 pp
31. Peréz CÁ, Montequín VR, Fernández FO, Balsera JV (2017) Integration of balanced scorecard (BSC), strategy map, and fuzzy analytic hierarchy process (FAHP) for a sustainability business framework: a case study of a spanish software factory in the financial sector. Sustainability 9:527
32. Porter ME (1997) Competitive strategy. Meas Bus Excel 1(2):12–17
33. Poveda-Bautista R, Baptista DC, García-Melón M (2012) Setting competitiveness indicators using BSC and ANP. Int J Prod Res 50(17):4738–4752
34. Prokop V, Stejskal J (2019) Different influence of cooperation and public funding on innovation activities within German industries. J Bus Econ Manag 20(2):384–397
35. Prokop V, Stejskal J, Hudec O (2019) Collaboration for innovation in small CEE countries. E a M: Ekonomie a Management 22(1):130–140
36. Rejeb HB, Morel-Guimaraes L, Boly V, Assielou NG (2008) Measuring innovation best practices: improvement of an innovation index integrating threshold and synergy effects. Technovation 28(12):838–854
37. Saunila M, Ukko J (2012) A conceptual framework for the measurement of innovation capability and its effects. Balt J Manag 7(4):355–375
38. Schentler P, Lindner F, Gleich R (2010) Innovation performance measurement. In: Gerybadze A, Hommel U, Reiners HW, Thomaschewski D (eds) Innovation and international corporate growth. Springer, Heidelberg, pp 299–317
39. Sohn MH, You T, Lee S-L, Lee H (2003) Corporate strategies, environmental forces, and performance measures: a weighting decision support system using the k-nearest neighbor technique. Expert Syst Appl 25:279–292
40. Sorayaei A, Abedi A, Khazaei R, Hossien Zadeh M, Agha Maleki SMSA (2014) An integrated approach to analyze strategy map using BSC–FUZZY AHP: a case study of dairy companies. Eur Online J Nat Soc Sci 2:1315–1322
41. Stejskal J, Hajek P (2019) Modelling collaboration and innovation in creative industries using fuzzy set qualitative comparative analysis. J Technol Transf 1–26
42. Woodside AG (2014) Embrace perform model: complexity theory, contrarian case analysis, and multiple realities. J Bus Res 67(12):2495–2503
43. Zhang Z (2016) Missing data imputation: focusing on single imputation. Ann Transl Med 4(1):1–8
44. Zizlavsky O (2016) Innovation scorecard: conceptual framework of innovation management control system. J Glob Bus Technol 12(2):10–27

Petr Hájek is an associate professor with the Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Czech Republic. He deals with the modelling of economic processes (especially in the field of public economics and public finance).

Jan Stejskal is an associate professor with the Institute of Economics, Faculty of Economics and Administration, University of Pardubice, Czech Republic. His domain is connection of the public economy in the regional scope and view. Especially, he analyses regional policy, tools of the local and regional economic development, and public services.

Michaela Kotková Stříteská is a professor assistant at Institute of Business Economics and Management, Faculty of Economics and Administration, University of Pardubice. She is specialist in BSC and strategic management field. She is a member of the research team titled Modelling knowledge spill-over effects in the context of regional and local development.

Viktor Prokop is a professor assistant at Institute of Economics Sciences, Faculty of Economics and Administration, University of Pardubice. The author is co-researcher of the grant project: Modelling knowledge spill-over effects in the context of regional and local development; and explores the issue of measuring the knowledge economy in his dissertation.

Innovation Environment in Europe—Efficiency Analysis Case Study

Viktor Prokop, Jan Stejskal, Petr Hájek, and Michaela Kotková Stříteská

Abstract Nowadays, measuring efficiency within countries' innovation environment seems to be incremental in the process of gaining competitive advantage. Therefore, this study is aimed to evaluate efficiency in patent creation within EU28 countries. We are using specialized tool for assessing the effectiveness, performance and productivity of comparable production units Data Envelopment Analysis and data from Eurostat. Moreover, we are analyzing countries' efficiency according to their innovation performance measured by European Commissions' European Innovation Scoreboard 2017. Results show that only 5 out of the 28 European countries are effectively using basic attributes of Innovation Environment (investment in science and research; human resources in science and technology; cooperation with external research and development firms). All of these countries belong to the group of Innovation Leaders. We also propose practical implications (for each country) on how to improve and how to change their inputs and outputs to become (more) efficient and provide information about countries that could be benchmark for less efficient countries.

Keywords Innovations · Entrepreneur environment · Efficiency analysis · Europe · Case study

V. Prokop · J. Stejskal (✉) · P. Hájek · M. Kotková Stříteská
Faculty of Economics and Administration, University of Pardubice, Pardubice, Czech Republic
e-mail: jan.stejskal@upce.cz

V. Prokop
e-mail: viktor.prokop@upce.cz

P. Hájek
e-mail: petr.hajek@upce.cz

M. Kotková Stříteská
e-mail: michaela.striteska@upce.cz

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_4

1 Introduction

There are dynamic changes in each country's economy and society. These are the unintended changes that are the result of the development of society and its globalization in the world. But it is also about the intended changes that are the result of applied public policies. For many years, developed countries have supported the expansion of knowledge and technological progress (innovation is the result), which are perceived as the basic engine of economic development [16]. Some authors have been tantalizing for diffusion, which, in their views, is a major determinant of the economic or social impact of innovation [6, 54].

A plethora of studies has confirmed the importance of innovation. Many other scholars have also addressed the environment in which innovation is emerging. The subject of the study was both efficiency and research of determinants that make up an innovative background or environment. These determinants become an important part of the production functions, which in a modern concept contain not only basic production factors but also knowledge variables or other soft variables [33].

Many studies (for example Drucker [19]; Rao [46]; Vila [53]; Andersson [4]; Meissner [36] and others); also analyze the activity and role of different subjects in this innovation environment. The familiar concept of triple-helix [21] and also the well-known endogenous growth models [12] show that public sector (state) and public subjects are currently one of the important players in the innovation environment.

It is becoming increasingly important to explore the innovation environment, its determinants and the roles of individual significant regional actors [39]. The knowledge about processes, mutual relations and relations, but also ways of financing or other public support can help us to define the suitable conditions of the efficiency of this innovative structures [2]. Particularly important is the transfer of knowledge from various advanced environments, where more advanced economic systems can be perceived as an important benchmark for less advanced countries (or those where some of the important elements of the environment fail or are not yet created).

This chapter is organized as follows. First, theoretical background is provided on the determinants of the innovation environment. Second, research methodology and underlying data are described. The next section provides the results of your own DEA models. The final section concludes this chapter and discusses the results and political implications.

2 Theoretical Background

As mentioned above, the innovation environment is dependent on the creation and diffusion of new knowledge [35]. Many studies have shown that knowledge generation is influenced by a number of determinants [23]. They are primarily regional variables, creativity, innovation, but also openness, freedom, etc. In more detail, it

has been found that the determinants of innovation need to be sought both in the microeconomic and macroeconomic environment of an innovative enterprise.

So many scientists have attempted to analyze the effectiveness of the innovation environment (in the region context) and regional innovation efficiency [39], as well as the performances of enterprises depending on the quality of the innovation environment (or selected determinants; [55]). They all agreed that it is necessary to perceive innovation in the context of the environment, to examine their effectiveness through the innovation approach system [1]. This is confirmed by the expanded concept of open innovation or open business models [49]. Nowadays, we can also capture transition from business ecosystem to innovation ecosystem that could be defined with different meanings and purposes (e.g. digital innovation ecosystem, hub ecosystems, open innovation ecosystem, platform-based ecosystem etc.) and which is understood as a set for the co-creation, or the jointly creation of value and innovation [17]. From these arguments there is a clear importance of innovation environment because of creating an open innovation environment that disclosure knowledge from the organizational to the inter-organizational level and opening up for external sources of knowledge flowing from the inter-organizational level to organizational level [45]. It allows environment and ecosystem stakeholders to assure the learning processes of the entire system.

For any efficiency analysis, as well as for the effectiveness of innovations or attributes of an innovative environment, we need to use the measurable indicators. Many studies (for example Hsu [30]; Ren [47] or Kim [32]) use innovation performance indicators such as patents, utility models, and sometimes revenue from innovative production. However, there are a number of critics, who do not perceive, for example, patent applications as sufficiently significant [9]. In practice, innovation takes place without their patenting, thanks to time considerations, financial demands and high dynamics in innovative change. Therefore, other ways to analyze innovative processes without these simplified indicators are being sought. Some come with their own indices, such as Innovation efficiency index by Guan and Chen [25].

In practice, qualitative studies based on empirical data are often used. They look for dependencies among the different variables that occur in the innovation environment and explore their impact on innovation performance. Similarly, these studies may only focus on selected industries or are suitable for macroeconomic studies [47]. Important frameworks for these researches are innovation systems, which in the regions represent the basis on which innovation activities are carried out between different actors. These are primarily regional innovation systems [11], sector innovation systems, or different types of innovation environment (milieu). These bases also allow you to analyze variables such as collaboration, the emergence and transfer of tacit knowledge or creativity [13, 52].

The core attributes of each innovation environment (see Fig. 1) operating altogether are:

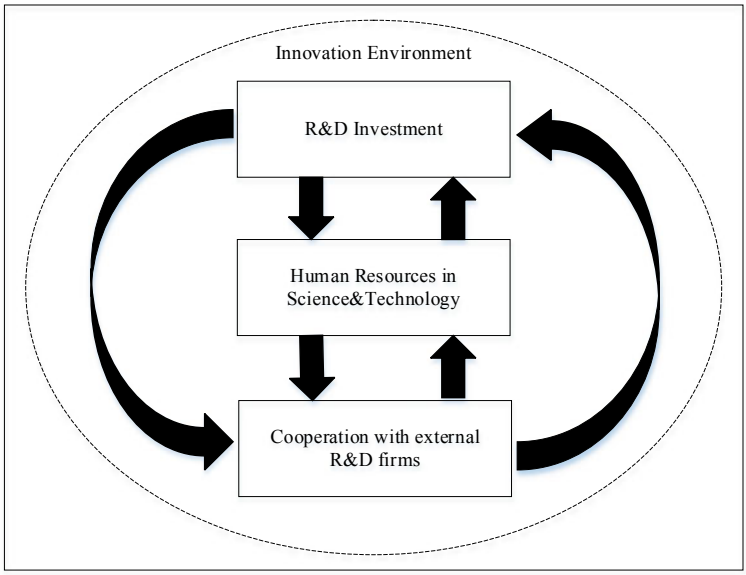


Fig. 1 Core attributes of innovation environment

- investments in science and research (including both private and public resources) for the acquisition of basic hard and soft infrastructure. There exists spatial dependence of R&D investments and efficiency while the regions with high efficiency of their R&D have spillover effect on the surrounding areas [29];
- HRST (human resources in science and technology)—internal knowledge resources—human capital, the holder of tacit and codified knowledge, innovation, creativity, etc. Innovation human resources represent the most dynamic resource among entrepreneurial enterprises (specifically high-tech) and play a very crucial role in the process of transforming innovation inputs into outputs within innovation environments [37];
- cooperation with external research and development firms—i.e. external knowledge sources) that could lead to higher innovation sales and more patents [51] and to creation of collaborative environment that increases the effects of abovementioned knowledge spillovers [26].

These attributes were also used in earlier publisher studies [10, 20, 28].

Even with these basic pillars of the innovation environment, it is possible to assume that there is no harmonious development in each country evenly. There is an innovation failure, a market failure, or a government failure to apply different public support policies. Moreover, these failures may give rise to adverse behavioral and relational consequences [34] and regularly endanger firms' overall competiveness (specifically in the case of new products and services failures [31]). Thus, macroeconomic efficiency analyzes are lacking, pointing to the different efficiency of creating

an innovative environment. Therefore, the aim of this chapter is to analyze the effectiveness of the basic attributes of the innovation environment in the 28 countries of the European Union.

In this research, we assume that there is a difference in efficiency in the creation of the innovation environment of individual countries. Similarly, we expect the highest efficiency to be achieved by countries that are considered innovative leaders according to the European Innovation Scoreboard.

3 Data and Research Method

For our analyses, we used parametric approach Data Envelopment Analysis (DEA) which is commonly used as a model specialized tool for assessing the effectiveness, performance and productivity of comparable production units (homogeneous units, e.g. countries of EU 28) based on the size of inputs and outputs. These units convert multiple inputs into outputs, meaning a set of units that produce the same or equivalent effects that are referred as the outputs of these units [41].

The principle of DEA models is that when evaluating the efficiency of a production unit it maximizes its efficiency level, assuming that the efficiency rate of all other DMUs cannot be higher than 1 (100%). The weights of all inputs and outputs must be greater than zero so that all the considered characteristics in the model are included (to see more [27]). The model can be built on the assumption of constant returns to scale (one unit of input generates one unit of output), when all DMUs are operating at optimal scale (CCR model). Rather unrealistic condition is solved by introducing variable returns to scale (VRS) considering all types of returns: increasing, constant or decreasing (BCC model).

For our cross-country analyses within the EU 28 countries, we used input-oriented VRS model operating with variable returns to scale. As a data source, we are using data from Eurostat databases. Number of researchers (e.g. Wang and Huang [54]) analyzed the optimal time delay between input and output variables. Griliches [24] empirically proved that there is no time delay with significant impact on the results of analyses. In this study, we are using 3 years' time delay between inputs (2014) and outputs (2017). Input and output variables are described below. In the next section, we compare countries efficiency and distinguish selected countries according to their innovation performance in 2017 measured by European Commission (Innovation Leaders, Strong Innovators, Moderate Innovators, and Modest Innovators).

4 Results

In the following part (see Table 1), the results of input-oriented VRS DEA model are shown for 28 European countries (Romania results are not displayed, there is lack of data). European countries that efficiently used selected inputs (see previous

Table 1 DEA efficiency results

	Country	Efficiency	Benchmarks	Inputs (2014)						Output (2017)	
				R&D exp. (% of GDP)		HRST (% of active pop.)		Ext. R&D coop. (% of firms)		Patents (per mil. pop.)	
				Orig.	*Adjust.*	*Orig.*	*Adjust.*	*Orig.*	*Adjust.*	*Orig.*	*Adjust.*
Innovation leaders	Sweden	**1.00000**	–	3.14	3.14	55.10	55.10	26.90	26.90	283.46	283.46
	Denmark-DN	**1.00000**	–	2.91	2.91	54	54.00	15.10	15.10	246.61	246.61
	Finland	0.88075	DE	3.17	2.79	55.60	48.97	50.60	20.62	235.68	235.68
	Netherlands-NL	**1.00000**	–	1.98	1.98	52.80	52.80	41.10	41.10	203.59	203.59
	UK	**1.00000**	–	1.66	1.66	54.60	54.60	–	–	82.62	82.62
	Germany-DE	**1.00000**	–	2.87	2.87	47	47.00	17.60	17.60	228.81	228.81
Strong innovators	Austria	0.98088	DE	3.08	2.88	48.30	47.38	25.70	18.03	231.35	231.35
	Luxembourg	0.87005	NL	1.26	1.10	64.50	37.78	24.90	19.72	93.94	93.94
	Belgium	0.77866	DE	2.39	1.86	51.10	39.79	35.40	13.38	145.83	145.83
	Ireland	0.65398	NL	1.55	1.01	53.10	34.73	27.70	13.73	77.64	77.64
	France	0.80598	DE	2.23	1.80	49.10	39.57	31.20	13.43	141.85	141.85
	Slovenia	0.69575	DE	2.37	0.94	43.70	30.40	42.70	5.81	55.30	55.30
Moderate innovators	Czech Republic	0.74399	DE	2	0.70	38.10	28.35	24.90	4.35	33.78	33.78
	Portugal	0.80106	DE	1.29	0.48	33	26.43	20.90	2.99	13.80	13.80
	Estonia	0.56759	DE	1.43	0.63	48.90	27.75	39.80	3.93	27.60	27.60
	Lithuania	0.55568	DE	1.03	0.41	46.50	25.84	10.30	2.57	7.57	7.57
	Spain	0.67574	DE	1.24	0.72	42.20	28.52	19.30	4.47	35.56	35.56
	Malta	0.67328	DN	0.72	0.48	39.50	26.59	4.40	2.96	14.40	14.40
	Italy	0.90466	DE	1.34	1.09	35	31.66	11.30	6.71	68.46	68.46

(continued)

Table 1 (continued)

	Country	Efficiency	Benchmarks	Inputs (2014)						Output (2017)	
				R&D exp. (% of GDP)		HRST (% of active pop.)		Ext. R&D coop. (% of firms)		Patents (per mil. pop.)	
				Orig.	*Adjust.*	*Orig.*	*Adjust.*	*Orig.*	*Adjust.*	*Orig.*	*Adjust.*
	Cyprus	0.83281	NL	0.51	0.42	48.80	26.36	21.50	3.48	10.62	10.62
	Slovakia	0.79285	DE	0.88	0.44	32.90	26.08	15.20	2.74	10.14	10.14
	Greece	0.73211	DE	0.83	0.42	35.40	25.92	19.90	2.62	8.38	8.38
	Hungary	0.74478	DE	1.35	0.55	36.30	27.04	13.80	3.42	20.08	20.08
	Latvia	0.64504	DE	0.69	0.45	40.70	26.25	18	2.92	11.41	11.41
	Poland	0.66446	DE	0.94	0.52	40.40	26.84	21.40	3.28	18.08	18.08
	Croatia	0.72934	DE	0.78	0.38	35.10	25.60	23.70	2.40	4.80	5.07
Modest innovators	Bulgaria	0.72316	DE	0.79	0.38	35.40	25.60	7.80	2.40	4.13	5.07
	Romania	–	–	–	–	–	–	–	–	–	–

Source Own according to Eurostat data

part—the core attributes of each innovation environment) in the process of Patent creation (output variable—Patent applications to the European Patent Office per million inhabitants) reached the rate of effectiveness 1,000. Countries that did not reach the rate of effectiveness 1,000 were not considered effective—less rate of effectiveness means less efficiency of the country.

In total, 5 out of the 28 EU countries (18%) were effective. All of these countries belong to the group of Innovation Leaders according to the European Commissions' European Innovation Scoreboard 2017. Only Finland was not efficient within the group of Innovation Leaders. On the other hand, countries that belong to the other groups according to their innovation performance were not efficient. Lithuania was evaluated as the least efficient country within analyzed EU 28 countries. Svagzdiene and Kuklyte [50] state that Lithuania, in comparison with the other EU 28 countries, payed minimum expenditure in the business sector and in the promotion of the economic policy did not consider innovation progress in a key priority for the driven agricultural sector and other industries.

The advantage of the DEA model is that it provides practical implications (for each country) on how to improve and how to change inputs and outputs to become (more) efficient [39]. Input-oriented models propose changes focusing primarily on input variables (or even minor changes on the output side). Table 1 therefore shows both original values (that each country reached) and adjusted values (provided by DEA) that show how the input (output) variables should be reduced/increased. Moreover, DEA also provide the information about countries that could be benchmark for other inefficient countries. Germany was proposed as benchmark country for other countries in 18 cases (66%). Germany represent the group of countries that are able to develop their innovation potential and create sufficient innovation environment and one of the global leaders in innovation and competitiveness, also in the context of the knowledge economy [42–44].

The results unambiguously confirm that there are significant differences in the innovation environment between innovation leaders and followers (in all categories). Almost all countries regarded as innovative leaders achieve maximum efficiency in creating a knowledge-based innovation environment. It is clear from the results that these countries achieve the optimal values of the individual variables and, even after the time delay, they have maximum efficiency. These results are in accordance e.g. with Prajogo [39] which pointed that dynamic efficient innovation environments strengthen the effect of innovation (specifically product innovation) on business performance. The significant role of core attributes within business environments and ecosystems (e.g. in terms of dynamism, innovation creation, growth and competitiveness) is indisputable. We can therefore see that our selected inputs represent contingency factors which affect the effectiveness of countries innovations (patent creation) in delivering countries performance. Moreover, Blazsek and Escribano [7] show that these countries (R&D and innovation leaders) have sustained future profitability and earn significant future excess returns, while followers only earn average returns. This could lead to the creation of dynamic (multiplier) spillover effects which may ultimately affect also following countries (R&D and innovation followers).

In the case of strong innovators, a lower level of effectiveness was found. In this category, it is possible to perceive differences between original and adjusted values and appropriate measures near the recommended benchmark. The result is that the smallest differences between the actual and the recommended value are in the number of patents. On the other hand, at microeconomic level, Prokop and Stejskal [40] show that firms from countries belonging to the group of strong innovators are able to effectively utilize the various determinants of innovation activities (e.g. financing from the EU, cooperation with clients or customers, cooperation with public research institutes, and expenditures in extramural R&D), to influence their innovation turnover. Therefore, there is a need to identify positive benchmarks for these countries (Germany and Netherlands) and follow these countries at macroeconomic level.

The worst results were found in the moderate and modest innovators group. It shows that there are almost three times less R&D expenditures in these countries. Important is that there has been a high level of cooperation and large numbers of HRSTs in these countries. However, given the results, it can be concluded that these elements do not have the required quality. R&D staff can not produce outputs that can be patented and do not represent significant global or at least European innovation. Ponsiglione et al. [38] also stated that innovation environments with similar industrial structures and characteristics can strongly differ from each other even in terms of innovation and competitive performance while this gap is more evident in the case of the so-called lagging regions (characterized by moderate and modest level of innovativeness). These lagging regions lacked solid interactions, network coordination, competences and skills as well [3]. This is, in our view, the cause of the low efficiency of innovation systems in these countries that are not able to increase their effectiveness and to reduce waste of innovation efforts. It could be also caused by a lack of exploiting new ideas by the firms and inertia caused by local systems [15].

5 Conclusions

An innovative environment is a vital complex element that affects most of the innovation processes in different organizations (both private and public). The state and its institutions or public organizations have the task of regulating the behavior of individuals and companies with the help of public policies. Their aim is, among other things, to create support schemes or financial schemes. Practice shows that there are significant differences between countries, namely in the quality of ensuring the basic attributes of the innovation environment.

The aim of this chapter was to analyze the effectiveness of the basic attributes of the innovation environment in EU countries 28. The basic attributes of the innovation environment were defined as inputs; the output variable was the number of patents. This was mainly a quantitative analysis.

The results are confirmed by the European Innovation Scoreboard, issued by the European Commission for the EU Member States. The countries that are considered

to be the European innovation leaders achieve the highest efficiency. Our analysis also showed the optimal values of individual indicators (input variables) that countries should achieve if they want to increase their efficiency.

These results can be used to define country-specific recommendations and their public policies. An essential recommendation for all countries with less efficiency is to ensure the basic attributes of the innovation environment—above all a skilled workforce with a strong knowledge base. Inefficient countries also should follow leaders' behavior which positively influences the ability of organizations to develop successful innovations (e.g. thanks to radical innovation, willingness to change, open innovation etc., [18]. However, these countries must take into account whether they have sufficient absorption capacity. Therefore, it is necessary to start from the bottom (e.g. changes in the education system, building a knowledge base, trust, etc.). Without these changes, innovation paradox or NIH and NSH (Not-Invented-Here, Not-Sold-Here) syndromes could occur. The latter syndromes (NIH, NSH) represent problems that can be rooted in corporate innovation culture and do not allow the company to adapt open innovation logic. It is connected with protective attitude towards external knowledge exploitation when employees affected by these syndromes feel that if the knowledge or technology cannot by exploited in own products/services, it should not be exploited at all by anyone else [14]. Frishammar et al. [22] state that understanding the firm's innovation culture is one of the most critical aspects to grasp when changing from a closed to a more open model of innovation and to mitigate these syndromes.

Ponsiglione et al. [38] state that the exploration capacity, the propensity to cooperation, and the endowed competencies of actors could be considered as key aspects in affecting innovation performance (specifically at regional level). Therefore, another recommendation is to increase attention to the results of co-operative links between subjects, to be dissatisfied with poor quality or low quality results. As the results of innovative followers show, some countries do not give a significant weight to quality. They report results (quantitatively high), but qualitatively these results do not contribute to the growth of the output variable (in this case the number of patents). Therefore, we propose implementing quality-orientated management (quality improvement methods—QIMs) that together with innovation represents central strategies for firms and poses significant managerial, organizational and technical challenges in the highly competitive international business world [8].

Thirdly, it is necessary to point the link between the quality of government and its components (e.g. rule of law, control of corruption, government effectiveness and government accountability) and the capacity of regions to innovate and to shape patenting [48] that play a key role mostly in countries suffering low innovative ability. It leads to the wide recognance that institutional factors influence innovative performances (specifically at regional and local level). Therefore, according to Arbolino et al. [5], we propose considering qualitative (socio-political aspects) and quantitative factors (managerial, policymaking and expenditure abilities). Zeng et al. [56] also state that firms should devote continuous efforts to maintain a solid quality system in place integrating a set of quality management (QM) practices (e.g. hard QM that pertains to the technical aspects of QM and soft QM that relates to the social/behavioral attributes) and corresponding performance measures.

The limitation of this study is the quality of the primary data entering the analysis and, of course, a certain limitation of the selected input indicators. Future research should be aimed at identifying the reasons for inefficiency of selected indicators of the innovation environment in individual European countries.

Acknowledgements This work was supported by a grant provided by the scientific research project of the Czech Sciences Foundation Grant No. 17-11795S.

References

1. Acs ZJ, Audretsch DB, Lehmann EE, Licht G (2017) National systems of innovation. J Technol Transf 42(5):997–1008
2. Adams R, Jeanrenaud S, Bessant J, Denyer D, Overy P (2016) Sustainability-oriented innovation: a systematic review. Int J Manage Rev 18(2):180–205
3. Anderson HJ, Stejskal J (2019) Diffusion efficiency of innovation among EU member states: a data envelopment analysis. Economies 7(2):34
4. Andersson U, Dasí À, Mudambi R, Pedersen T (2016) Technology, innovation and knowledge: the importance of ideas and international connectivity. J World Bus 51(1):153–162
5. Arbolino R, Boffardi R, De Simone L (2018) Which are the factors influencing innovation performances? Evidence from Italian Cohesion Policy. Social Indicators Research, pp 1–27
6. Berry FS, Berry WD (2018) Innovation and diffusion models in policy research. In: Theories of the policy process. Routledge, pp 263–308
7. Blazsek S, Escribano A (2016) Patent propensity, R&D and market competition: dynamic spillovers of innovation leaders and followers. J Econometrics 191(1):145–163
8. Bourke J, Roper S (2017) Innovation, quality management and learning: short-term and longer-term effects. Res Policy 46(8):1505–1518
9. Brem A, Nylund PA, Schuster G (2016) Innovation and de facto standardization: the influence of dominant design on innovative performance, radical innovation, and process innovation. Technovation 50:79–88
10. Caragliu A, Nijkamp P (2012) The impact of regional absorptive capacity on spatial knowledge spillovers: the Cohen and Levinthal model revisited. Appl Econ 44(11):1363–1374
11. Cooke P, Uranga MG, Etxebarria G (1997) Regional innovation systems: institutional and organisational dimensions. Res Policy 26(4–5):475–491
12. Cozzi G (2017) Endogenous growth, semi-endogenous growth… or both? A simple hybrid model. Econ Lett 154:28–30
13. Cumbers A, Mackinnon D, Chapman K (2003) Innovation, collaboration, and learning in regional clusters: a study of SMEs in the Aberdeen oil complex. Environ Plann A 35(9):1689–1706
14. Dąbrowska J, Teplov R, Podmetina D, Albats E, Lopez-Vega H (2017) Ready or not? Organizational capabilities of open innovation adopters and non-adopters. In: ISPIM conference proceedings. The International Society for Professional Innovation Management (ISPIM), pp 1–21
15. Das P, Verburg R, Verbraeck A, Bonebakker L (2018) Barriers to innovation within large financial services firms: an in-depth study into disruptive and radical innovation projects at a bank. Eur J Innov Manage 21(1):96–112
16. David PA (1992) Knowledge, property, and the system dynamics of technological change. World Bank Econ Rev 6(suppl_1):215–248
17. de Vasconcelos Gomes LA, Facin ALF, Salerno MS, Ikenami RK (2018) Unpacking the innovation ecosystem construct: evolution, gaps and trends. Technol Forecast Soc Chang 136:30–48

18. Domínguez-Escrig E, Mallén-Broch FF, Lapiedra-Alcamí R, Chiva-Gómez R (2018) The influence of leaders' stewardship behavior on innovation success: the mediating effect of radical innovation. J Bus Ethics 1–14
19. Drucker PF (1985) The discipline of innovation. Harvard Bus Rev 63(3):67–72
20. Erken H, Kleijn M (2010) Location factors of international R&D activities: an econometric approach. Econ Innov New Technol 19(3):203–232
21. Etzkowitz H, Ranga M (2015) Triple Helix systems: an analytical framework for innovation policy and practice in the Knowledge Society. In: Entrepreneurship and knowledge exchange. Routledge, pp 117–158
22. Frishammar J, Richtnér A, Brattström A, Magnusson M, Björk J (2019) Opportunities and challenges in the new innovation landscape: implications for innovation auditing and innovation management. Eur Manag J 37(2):151–164
23. González-Pernía JL, Kuechle G, Peña-Legazkue I (2013) An assessment of the determinants of university technology transfer. Econ Dev Q 27(1):6–17
24. Griliches Z (1998) Patent statistics as economic indicators: a survey. In: R&D and productivity: the econometric evidence. University of Chicago Press, pp 287–343
25. Guan J, Chen K (2010) Measuring the innovation production process: a cross-region empirical study of China's high-tech innovations. Technovation 30(5–6):348–358
26. Hajek P, Stejskal J (2018) R&D cooperation and knowledge spillover effects for sustainable business innovation in the chemical industry. Sustainability 10(4):1064
27. Halaskova M, Halaskova R, Prokop V (2018) Evaluation of efficiency in selected areas of public services in European union countries. Sustainability 10(12):4592
28. Hazır CS, LeSage J, Autant-Bernard C (2018) The role of R&D collaboration networks on regional knowledge creation: evidence from information and communication technologies. Pap Reg Sci 97(3):549–567
29. He B, Wang J, Wang J, Wang K (2018) The impact of government competition on regional R&D efficiency: does legal environment matter in China's innovation system? Sustainability 10(12):4401
30. Hsu CW, Lien YC, Chen H (2015) R&D internationalization and innovation performance. Int Bus Rev 24(2):187–195
31. Joachim V, Spieth P, Heidenreich S (2018) Active innovation resistance: an empirical study on functional and psychological barriers to innovation adoption in different contexts. Ind Mark Manage 71:95–107
32. Kim B, Kim E, Miller DJ, Mahoney JT (2016) The impact of the timing of patents on innovation performance. Res Policy 45(4):914–928
33. Klímová V, Žítek V, Králová M (2019) How public R&D support affects research activity of enterprises: evidence from the Czech Republic. J Knowl Econ. https://doi.org/10.1007/s13132-019-0580-2
34. Liao S, Chou CY, Lin TH (2015) Adverse behavioral and relational consequences of service innovation failure. J Bus Res 68(4):834–839
35. Maskell P (2001) Knowledge creation and diffusion in geographic clusters. Int J Innov Manag 5(02):213–237
36. Meissner D, Polt W, Vonortas NS (2017) Towards a broad understanding of innovation and its importance for innovation policy. J Technol Transf 42(5):1184–1211
37. Pan X, Zhang J, Song M, Ai B (2018) Innovation resources integration pattern in high-tech entrepreneurial enterprises. Int Entrepreneurship Manage J 14(1):51–66
38. Ponsiglione C, Quinto I, Zollo G (2018) Regional innovation systems as complex adaptive systems: the case of lagging European regions. Sustainability 10(8):2862
39. Prajogo DI (2016) The strategic fit between innovation strategies and business environment in delivering business performance. Int J Prod Econ 171:241–249
40. Prokop V, Stejskal J (2017) Different approaches to managing innovation activities: an analysis of strong, moderate, and modest innovators. Inžinerine Ekonomika/Eng Econ 28(1)
41. Prokop V, Stejskal J (2017) Effectiveness of knowledge economy determinants: case of selected EU members. In: European conference on knowledge management. Academic Conferences International Limited, pp 825–832

42. Prokop V, Stejskal J (2019) Different influence of cooperation and public funding on innovation activities within German industries. J Bus Econ Manage 20(2):384–397
43. Prokop V, Stejskal J, Hudec O (2019) Collaboration for innovation in small CEE countries. E + M Ekonomie a Management 22(1):130–144
44. Prokop V, Stejskal J, Hajek P (2018) Effectiveness of selected knowledge-based determinants in macroeconomics development of EU 28 economies. In: Finance and economics readings. Springer, Singapore, pp. 69–83
45. Radziwon A, Bogers M (2019) Open innovation in SMEs: exploring inter-organizational relationships in an ecosystem. Technol Forecast Soc Chang 146:573–587
46. Rao S, Ahmad A, Horsman W, Kaptein-Russell P (2001) The importance of innovation for productivity. Int Prod Monit 2(spring):11–18
47. Ren S, Eisingerich AB, Tsai HT (2015) Search scope and innovation performance of emerging-market firms. J Bus Res 68(1):102–108
48. Rodríguez-Pose A, Di Cataldo M (2014) Quality of government and innovative performance in the regions of Europe. J Econ Geogr 15(4):673–706
49. Stejskal J, Mikušová Meričková B, Prokop V (2016) The cooperation between enterprises: significant part of the innovation process: a case study of the czech machinery industry. E + M Ekonomie a Management 19(3):110–122
50. Svagzdiene B, Kuklyte J (2016) The analysis of factors which have impact for summary innovation index in Germany, Estonia and Lithuania. Transform Bus Econ 15
51. Szücs F (2018) Research subsidies, industry–university cooperation and innovation. Res Policy 47(7):1256–1266
52. Tucci CL, Chesbrough H, Piller F, West J (2016) When do firms undertake open, collaborative activities? Introduction to the special section on open innovation and open business models. Ind Corp Change 25(2):283–288
53. Vila N, Kuster I (2007) The importance of innovation in international textile firms. Eur J Mark 41(1/2):17–36
54. Wang EC, Huang W (2007) Relative efficiency of R&D activities: a cross-country study accounting for environmental factors in the DEA approach. Res Policy 36(2):260–273
55. Zanello G, Fu X, Mohnen P, Ventresca M (2016) The creation and diffusion of innovation in developing countries: a systematic literature review. J Econ Surv 30(5):884–912
56. Zeng J, Phan CA, Matsui Y (2015) The impact of hard and soft quality management on quality and innovation performance: an empirical study. Int J Prod Econ 162:216–226

Viktor Prokop is a professor assistant at Institute of Economics Sciences, Faculty of Economics and Administration, University of Pardubice. The author is co-researcher of the grant project: modelling knowledge spill-over effects in the context of regional and local development; and explores the issue of measuring the knowledge economy in his dissertation.

Jan Stejskal is an associate professor with the Institute of Economics, Faculty of Economics and Administration, University of Pardubice, Czech Republic. His domain is connection of the public economy in the regional scope and view. Especially, he analyses regional policy, tools of the local and regional economic development, and public services.

Petr Hájek is an associate professor with the Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Czech Republic. He deals with the modelling of economic processes (especially in the field of public economics and public finance).

Michaela Kotkova Striteska is a professor assistant at Institute of Business Economics and Management, Faculty of Economics and Administration, University of Pardubice. She is specialist in BSC and strategic management field. She is a member of the research team titled Modelling knowledge spill-over effects in the context of regional and local development.

Application of Artificial Intelligence in Modeling a Textile Finishing Process

Zhenglei He, Kim Phuc Tran, Sébastien Thomassey, Xianyi Zeng and Changhai Yi

Abstract Textile products with faded effect are increasingly popular nowadays. Ozonation is a promising finishing process treatment for obtaining such effect in the textile industry. The interdependent effect of the factors in this process on the products' quality is not clearly known and barely studied. To address this issue, the attempt of modeling this textile finishing process by the application of several artificial intelligent techniques is conducted. The complex factors and effects of color fading ozonation on dyed textile are investigated in this study through process modeling the inputs of pH, temperature, water pick-up, time (of process) and original color (of textile) with the outputs of color performance (K/S, L^*, a^*, b^* values) of treated samples. Artificial Intelligence techniques included ELM, SVR and RF were used respectively. The results revealed that RF and SVR perform better than ELM in stably predicting a certain single output. Although both RF and SVR showed their potential applicability, SVR is more recommended in this study due to its balancer predicting performance and less training time cost.

Z. He (✉) · S. Thomassey
GEMTEX – Laboratoire de Génie et Matériaux Textiles, ENSAIT, 59000 Lille, France
e-mail: zhenglei.he@ensait.fr

S. Thomassey
e-mail: sebastien.thomassey@ensait.fr

K. P. Tran · X. Zeng
ENSAIT & GEMTEX, Roubaix, France
e-mail: kim-phuc.tran@ensait.fr

X. Zeng
e-mail: xianyi.zeng@ensait.fr

C. Yi
National Local Joint Engineering Laboratory for Advanced Textile Processing and Clean Production, Wuhan Textile University, Wuhan 430200, China
e-mail: 751067874@qq.com

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_5

Abbreviations

ELM	Extreme Learning Machine
SVR	Support Vector Regression
SVM	Support vector Machine
RF	Random Forest
ANN	Artificial Neural Network
SLFNs	Single-Layer-Feedforward-Neural-Networks
RB-RN	Reactive blue FL-RN
RR-2BL	Reactive red FL-2BL
RY-2RN	Reactive yellow FL-2RN
MAE	Mean absolute error
RMSE	Root mean square error
R	Correlation coefficient
MRAE	Mean relative absolute error
LOO	Leave one out
MRF	Multivariate Random Forest
W	Input weight matrix
β	Output weights
ε	Threshold
ξ_i	Slack variables, upper constraints on the outputs of the system
ξ_i^*	Slack variables, lower constraints on the outputs of the system
L	Lagrangian
$\eta_i, \eta_i^*, \alpha_i^*, \alpha_i^*$	Lagrange multipliers
γ, λ	Positive regularized parameters controlling the bias-variance trade-off in SVM
p	Parameter of RBF that sets the spread of the kernel
ntree	The number of trees in the forest
minleaf	Minimum number of samples in the leaf node
mtry	Randomly selected features considered for a split in each regression tree node

1 Introduction

1.1 Color Fading Ozonation: A Textile Finishing Process

In recent years, textile products with faded effect, worn look and vintage style, are attracting a growing number of young customers' attention and have gained a considerable share of the fashion market [1]. However, the faded effect of these products is achieved by a textile finishing process consisted by a large number of chemical treatments (e.g. bleaching using hydrogen peroxide or chlorine, washing using stone/permanganate). The traditional use of these chemical treatments is not

only highly water- and power-consuming, but also release a wide range of toxic substances to the environment. Over the past few decades, the rising concern of people on the environment issues has leaded to the rapid development of alternative sustainable approaches, where ozonation is the most important one of its in textile finishing.

Ozone is an excellent gaseous oxidant with an environmentally friendly nature that can be rapidly decomposed into O_2 after its application without emitting additional pollution. It is able to react with a large number of organic and inorganic substances in water due to a series of intermediates or by-products such as hydroxyl radicals (which reacts with no selectivity) may be generated in the reaction between ozone and water [2]. More significantly, ozone could be applied directly in the form of gas without a water bath that react with the targets (with certain water content directly and consequently dramatically decrease the water consumption in the sector. Meanwhile, approximate performance of conventional treatments on color fading effect can be obtained in the ozonation with less damage and other negative influence on target product materials [3]. Therefore, it is regarded as a perfect alternative to traditional oxidizing agents and bleaching agents and has been applied to a wide range of textile related domains such as wastewater treatment, dyeing, paper-making, fiber modification. The studies regarding color fading dyed textile using ozone, instead of the conventional processes, have been increasingly reported in these years by taking advantage of the application of ozone [4].

The decolorization of dyes in ozonation, in short, could be attributed to the simultaneous oxidation of direct ozone and (but more of) indirect free radicals (which is generated from the decomposition of ozone) with the unsaturated organic compounds, chromophoric organic system, e.g. the chromophore groups of azo. However, the real situation of the actions in the application of ozonation for color fading dyed textile is very complicated that can hardly be concluded in couples of sentences. Color fading ozonation of textiles is affected by many interdependent different factors ranging from the properties of textile material to the setting of color fading process [5]. How these factors affect the color fading process separately is known, while to understand their overall impacts simultaneously, the complex and nonlinear relationship between the factors of material properties as well as technical parameters of ozonation and color fading effects must be taken into consideration. In previous literatures, the simultaneous effects of multiple factors on color fading ozonation of textiles have been barely systematically investigated. This is because the factors have an extreme nonlinear and hardly-understood relation as well as unclear effects on the target product properties, analytical or mathematical expression in terms of models relied on chemical or physical laws with certain simplified assumptions for understanding the mechanism is limited in this regard. To address this large number input and output parameter issue, using artificial intelligent techniques that can learn from data would be more effective and applicable.

1.2 Artificial Intelligent Techniques for Modeling Textile Process

As the computing power has increasingly promoted in these years, the related intelligent modeling techniques are well developed, while as the correlation between wanted inputs and outputs in textile manufacturing process are hardly to be characterized, the applicable techniques mostly were concentrated in multiple linear or nonlinear regression models, ANN etc. ANN is a widely used artificial intelligence approach in the textile sector. It is a method inspired by the bionic simulation of human brain that interconnected numerous neurons in different hidden layers to process the complex information of specific input-output relation [6]. In particular, ELM is a novel algorithm for SLFNs, which randomly chooses W and analytically determines β of SLFNs. ELM tends to acquire a good generalization performance at extremely fast learning speed [7]. Sun et al. have successfully applied ELM to forecast sales behavior in the fashion retailing, and the experimental results in their study demonstrated that this ELM model performed better than backpropagation neural networks based methods [8].

SVM is also a popular machine learning tool based on artificial intelligence for classification and regression based on statistic learning theory, first identified by Vladimir Vapnik and his colleagues in 1992 [9]. SVR is the most common application form of SVM and a typical feature of it is that SVR minimizes the generalized error bound instead of minimizing the observed training error so as to achieve generalized performance. And it only relies on a subset of the training data due to the cost function for building the model neglects any training data that is close (within ε) to the model prediction [10]. The excellent use of SVR in textile industry has been issued for predicting yarn properties [11, 12], PU-coated cotton fabrics qualities [13] and wool knitwear pilling propensity [14], which have shown the potential of SVR in the application of textile process modeling.

RF is another famous artificial intelligence based model technique that composed of a weighted combination of multiple regression trees. It constructs each tree using a different bootstrap sample of the data, and different from the decision tree splitting each node using the best split among all variables, RF using the best among a subset of predictors randomly chosen at that node [15]. In general, combining multiple regression trees increases predictive performance. It accurately predicts by taking advantage of the interaction of variables and the evaluation of the significance of each variable [16]. Kwon et al. [17] developed a surface defect detection method based on RF to inspect the fabric surface. Venkatraman and Alsberg [18] predicted the important photovoltaic properties of phenothiazine dyes using RF which paves the way for rapid screening of new potential dyes and computer-aided materials design.

1.3 Modeling Color Fading Ozonation of Reactive-Dyed Cotton

Cotton is the most vital material in textile industry, and reactive dyes take the dominant position in cotton dyeing as it is easy to achieve optimum dyeing performance on cotton with environmentally-friendly advantages without high cost and complicated processes. These have benefited the reactive-dyed cotton to be one of the most important textile products in the fashion market. Color is the most important property of a textile product which finally determined by the finishing process. According to Kubelka-Munk theory [19], it is known that K/S value can indicate the color depth of textile products. While L^*, a^*, b^* values (or CIELab), an international standard widely used for color measurements, is capable of illustrating the color variation of textile samples. Among these colorimetric values, L^* (ranges from 0 to 100) is the lightness component, whereas a^* and b^* are chromatic components and demonstrate the color variation from green ($\triangle a^* < 0$) to red ($\triangle a^* > 0$) and blue ($\triangle b^* < 0$) to yellow ($\triangle b^* > 0$) respectively by a series of numbers from -120 to 120. Normally, the color of the final textile product agreeing with specific K/S and L^*, a^*, b^* values is in the acceptable tolerance of the consumer. Therefore, the K/S and $L^*a^*b^*$ values could be used to characterize the color variation of the color fading ozonation on reactive-dyed cottons.

An attempt is made for modeling color fading ozonation, a textile finishing process, in order to predict the color properties of ozone faded reactive-dyed cotton using different artificial intelligent techniques. ELM, SVR and RF model were constructed with corresponding optimization process to comparatively find the potential applicability of them in predicting the color performance of the reactive-dyed cotton in a textile finishing process named color fading ozonation. Part of this work can also be found in Ref. [20].

2 Experimental

2.1 Material

Desized grey cotton fabrics ($^3/_1$ twill; 325.7 g/m^2; supplied by Shunfu, Hubei, China) were dyed by three bifunctional fluorotriazine azo reactive dyes of RB-RN, RR-2BL and RY-2RN (provided by Color Root, Hubei, China; commercial quality, purity of dyes: 92%) respectively. Chemical material such as sodium hydroxide, hydrogen chloride, sodium metasilicate nonahydrate, 30% hydrogen peroxide, sodium sulfate, sodium carbonate (analytical grade, supplied by Sinopharm Limited, China) and OP-20 (polyoxyethylene octylphenol ether, a nonionic surfactant, chemical pure, supplied by Tianjin Guangfu, China) were used in this study.

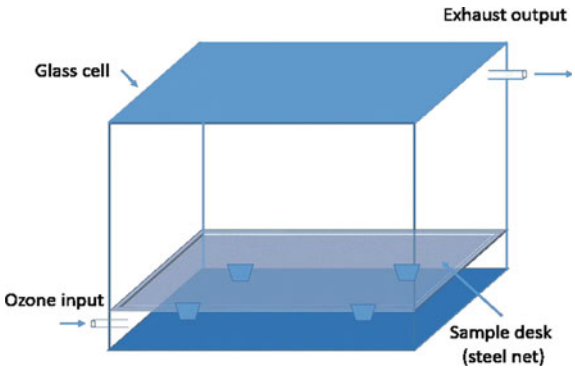


Fig. 1 The reactor setup of ozonation

2.2 Apparatus

Ozone employed in this work was generated by a corona discharge ozone generator, CF-G50 (Guolin, China), that fed by pure and compressed dry oxygen (≥99.9%, 1 Mpa, 12 L/min) from oxygen cylinder. Ozone was flow to the reactor (made of glass, the structure is exhibited in Fig. 1), and in each single color fading ozonation experiment, samples would be distributed evenly on the sample desk (made of air-permeable steel net). Ozone was imported with a gas flow of 2 L/min and a dosage of 137 ± 3 mg/L min (tested by UV meter NS-xmd614, Naishi, China) throughout the treatment. The exhaust from reactor would be collected and decomposed by a heater (≥230 °C) before evacuating to the atmosphere.

2.3 Methods

2.3.1 Pretreatment and Dyeing of Cotton Fabrics

The cotton fabric was scoured with 8 g/L sodium hydroxide, 3 g/L OP-20 and 5 g/L sodium metasilicate nonahydrate at 100 °C for 15 min at a liquor ratio of 20:1 and then was bleached with 8 g/L hydrogen peroxide, 3 g/L OP-20 and 5 g/L sodium metasilicate nonahydrate at 90 °C for 15 min at a liquor ratio of 20:1. Afterward, rinsing the fabrics thoroughly before it was dyeing with 3% o.w.f (on weight of fabric) dyestuff (RB-RN, RR-2BL and RY-RN respectively), 70 g/L sodium sulfate, 20 g/L sodium carbonate and 2 g/l OP-20 at a liquor ratio of 20:1 based on the profile displayed in Fig. 2.

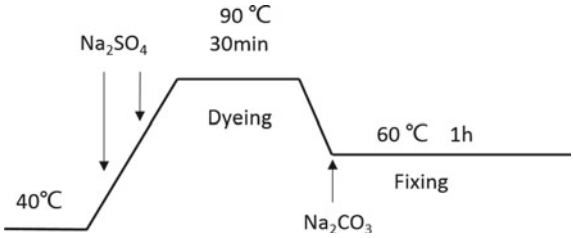


Fig. 2 Dyeing profile of the reactive dyes

2.3.2 Ozonation Process

Three dyed cottons in different colors were treated respectively by the color fading ozonation following the steps: wetting the fabrics by deionized water (pH = 7, or using sodium hydroxide and hydrogen chloride respectively when specific pH is required) to obtain certain pick-up water content. After ozone treating, samples were rinsed by deionized water before naturally drying up.

Ozonation at different pH (1, 4, 7, 10, 13), temperature (0, 20, 40, 60, 80 °C) with variable pick-ups (water content of sample, 0, 75%, 150%)) for different treating time (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60 min) were investigated on the three dyed cotton fabrics (blue, red, yellow, 612 fabrics samples in total) respectively. Besides of pH which was set up depending on the method mentioned above (using sodium hydroxide and hydrogen chloride respectively in the water pick-up step), the temperature of ozonation was controlled by a water bath around the reactor (including the inlet tubes), and the pick-up of sample was calculated by the Eq. (1).

$$Pickup\,(\%) = \frac{W_s - W_0}{W_0} \times 100\%, \tag{1}$$

where W_s and W_0 are the weights of the wet pickup sample and corresponding non-wetting original sample respectively.

2.3.3 Analytical

Colorimetric values of K/S and L^*, a^*, b^* were tested by Datacolor 110 spectrophotometer (Datacolor, USA) from taking the average of four measurements located on different parts of two sides of each sample, within a relative error of 0.3. All of the samples were conditioned at 21 $\pm$ 1 °C with moisture of 65 $\pm$ 2% over 24 h before each ozonation process and the test experiment.

3 Algorithms of Intelligent Techniques and Structure for Modeling

3.1 Extreme Learning Machine

ELM is an algorithm of SLFNs randomly chooses W and analytically determines β. Taking K hidden nodes SLFNs as an example, using activation function $f(x) = (f_1(x), f_2(x), ..., f_k(x))$ to learn N samples (X_i, Y_i), where $X_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \in R_n$ and $Y_i = [y_{i1}, y_{i2}, ..., y_{in}]^T \in R_m$. The ideal approximation of the SLFNs to these samples is zero error, which turns out that

$$\sum_{j=1}^{N} \|\hat{Y}_j - Y_j\| = 0 \tag{2}$$

where $\hat{Y}$ is the actual output value of SLFNs. Taking the weights W, β and bias b into consideration, we have

$$\sum_{i=1}^{K} \beta \cdot f_i(W_i \cdot X_j + b_i) = Y_j, j = 1, ..., N \tag{3}$$

where $W_i = [w_{i1}, w_{i2}, ..., w_{im}]^T$ and $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T, i = 1, ...K$ are the weight vector for inputs and activated nodes respectively. b_i is the threshold of i_{th} hidden node. The compact expression of Eq. (3) terms of vectorization could be

$$H\beta = Y \tag{4}$$

where $H(W_1, ..., W_j, b_j, ..., b_j, X_1, ..., X_i) = f(W_j \cdot X_i + b_j)$ $(i = 1, ..., N$ and $j = 1, ..., K)$ is the hidden layer output matrix of the neural network, the j_{th} column of it is the j_{th} hidden node output in regard to the inputs of $X_1, ..., X_i$. While $\beta = [\beta_1, \beta_2, ..., \beta_K]^T$ and $Y = [Y_1, Y_2, ..., Y_N]^T$ are the matrix of output weights and targets respectively. As the input weights W are randomly chosen, as well as the biases b in the ELM algorithm, the output weights β which connect the hidden layer and output layer could be simply determined by finding the least-square solution to the given linear system. According to [21], the smallest norm least-squares solution of the linear system (4) among all the solutions is

$$\hat{\beta} = H^{\dagger}Y \tag{5}$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse of the matrix H [22]. In this study, A multi-output ELM regression function developed by Huang's group was used in this study with an optimal trial of the varied activation functions (i.e. Sigmoid, Sine,

and Hardlim) given in Eqs. (6)–(8) and the number of hidden nodes (from 1 to 200) in the use of ELM.

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{6}$$

$$Sine(x) = sin(x) \tag{7}$$

$$Hardlim(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0 \end{cases} \tag{8}$$

3.2 Support Vector Machine

Compared with neural networks, SVR assures more generalization on the foundation of structural risk minimization, and generally performs better with less training samples. When we have training data (x_l, y_l), ..., $(x_l, y_l) \subset \mathbb{R}^n \times \mathbb{R}$ for the SVR model, the targeted function $g(x)$ should be as plat as possible and has ε deviation in maximum from the actual targets yi for all the training data in the form of:

$$g(x) = <w, x> + b; \quad w \in \mathbb{R}^n, b \in \mathbb{R} \tag{9}$$

where x is the n-dimensional input vectors, w is the weight vector and b is the bias term. Flatness in (9) means small w, and the way achieving it is recommended to minimize the Euclidean norm, i.e. $\frac{1}{2}\|w\|^2$ [23], which turns out to a convex optimization problem:

$$minimize \quad \frac{1}{2}\|w\|^2 \tag{10}$$

$$subject\ to \begin{cases} y_i - <w, x_i> - b \leq \varepsilon \\ <w, x_i> + b - y_i \leq \varepsilon \end{cases} \tag{11}$$

This is a feasible optimization problem when the function $g(x)$ actually exists and approximates all pairs (x_i, y_i) with ε precision, and ξ_i, ξ_i^* were introduced to deal with the otherwise infeasible constraints of it [9],

$$minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{12}$$

$$subject\ to \begin{cases} y_i - < w, x_i > -b \leq \varepsilon + \xi_i \\ < w, x_i > +b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{13}$$

where C is a constant greater than 0, determines the trade-offs of $\frac{1}{2}\|w\|^2$ and the sum of permitted errors. It is found that dual formulation makes it easily to solve this optimization problem [24], a standard dualization method utilizing Lagrange multipliers has been proposed:

$$\begin{aligned} L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - <w, x_i> + b) \\ - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - <w, x_i> - b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{14}$$

where $\eta_i, \eta_i^*, \alpha_i^*, \alpha_i^*$ have to satisfy positivity constraints of ≥ 0. The partial derivatives of L with respect to the variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality.

$$\partial_b L = \sum_{1=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{15}$$

$$\partial_w L = w - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i = 0 \tag{16}$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{17}$$

here $\xi_i^{(*)}$, $\alpha_i^{(*)}$ and $\eta_I^{(*)}$ refers to ξ_i and ξ_i^*, α_i and α_i^*, η_i and η_i^*, respectively. Substituting (15)–(17) into (14), the dual optimization problem is given by

$$maximize \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)<x_i, x_j> \\ -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{18}$$

$$subject\ to \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \quad and \quad \alpha_i, \alpha_i^* \in [0, C] \tag{19}$$

As the dual variables η_i, η_i^* can be reformulated on the basis of (18), (19) as $\eta_i^{(*)} = C - \alpha_i^{(*)}$, (16) turns to

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i, \quad thus \quad g(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)<x_i, x> + b \tag{20}$$

This is a so-called *Support Vector expansion*. In SVM training algorithm, the next necessary step is to make it nonlinearly, which was suggested to be achieved by a

mapping $\phi(x)$ from $\mathbb{R}^n$ to a higher dimensional feature space using kernel function $K(x, x_i) = <\phi(x_i), \phi(x)>$, therefore (20) becomes

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\phi(x_i), \tag{21}$$

$$g(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{22}$$

It is different from the linear case as w means the flatness is no longer explicitly given. In this nonlinear case, the optimization problem refers to finding the flattest function in feature space, rather than in input space. The standard SVR is

$$g(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{23}$$

where N (should be less than the total number of input-output pairs) is the number of input data having nonzero values of $\alpha_i^{(*)}$. The kernel function $k(x_i, x)$ corresponds to a linear dot product of the nonlinear mapping. As we are disposing of a case of the process modeling containing multiple outputs, we applied a multi-output least-squares support vector regression (MLS-SVR) toolbox developed by Xu et al. [25]. Particularly in this study with an optimal trail on kernel functions of:

$$\textit{Linear:}\quad K(x, x_i) = x^T x_i + C \tag{24}$$

$$\textit{Sigmoid:}\quad K(x, x_i) = tanh(\alpha x^T x_i + C) \tag{25}$$

$$\textit{Polynomial:}\quad K(x, x_i) = <x, x_i>^p \tag{26}$$

$$\textit{Radial basis function:}\quad K(x, x_i) = e^{-\frac{\|x - x_i\|^2}{2\sigma^2}} \tag{27}$$

$$\textit{Exponential radial basis function:}\quad K(x, x_i) = e^{-\frac{\|x - x_i\|}{2\sigma^2}} \tag{28}$$

There is also an optimization process (LOO) of the parameters of γ, λ and p in the toolbox from where $\gamma \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$, $\gamma \in \{2^{-10}, 2^{-8}, ..., 2^{10}\}$ and $p \in \{2^{-15}, 2^{-13}, ..., 2...3\}$ $(=\frac{1}{2\sigma^2})$ [25].

3.3 Random Forest

RF is an ensemble-learning algorithm depending on the bagging method that combines multiple independently-constructed decision tree predictors to classify or predict certain variables [16]. In RF, successive trees do not rely on earlier trees; they are independently using a bootstrap sample of the dataset, and therefore a simple unweighted average over the collection of grown trees $h(x, \Theta_k)$ would be taken for prediction in the end.

$$\bar{h}(X) + \frac{1}{K}\sum_{k=1}^{K} h(X, \Theta_k) \tag{29}$$

where $k = 1, ..., K$ is the number of trees, x represents the observed input vector, Θ is an independent identically distributed random vector that the tree predictor takes on numerical values. RF algorithm starts from randomly drawing *ntree* bootstrap samples from the original data with replacement. And then grow the certain number of regression trees in accordance with the bootstrap samples. In each node of regression tree, a number of the best split (*mtry*) randomly selected from all variables are considered for binary partitioning. The selection of the feature for node splitting from a random set of features decreases the correlation between different trees and thus the average prediction of multiple regression trees is expected to have lower variance than individual regression trees [26]. Regression tree hierarchically gives specific restriction or condition and it grows from the root node to the leaf node by splitting the data into partitions or branches according to the lowest Gini index:

$$I_G(t_{X(X_i)}) = 1 - \sum_{J=1}^{M} f(t_{X(X_i)}, j)^2 \tag{30}$$

where $f(t_{X(X_i)}, j)$ is the proportion of samples with the value x_i belonging to leaf j as node t [27]. In present study, MRF developed by Raziur Rahman et al. [5] was employed with an optimal topology of three parameters in terms of *ntree*, *minleaf* and *mtry*.

3.4 Modeling Structure

In this work, constructed model is expected to be capable of predicting (or outputting) the color qualities of ozone treated samples in terms of K/S and L^*, a^*, b^* values by giving 5 variables including not only the specific color of treated fabric but also the process parameters of pH, temperature, pick-up and treating time. In other words, the anticipated model of color fading ozonation on reactive-dyed cotton realizes the complex and unclear relationship of color fading ozonation parameters and its effectiveness on reactive-dyed cotton fabric in certain respects.

Particularly, taking the real samples used in the ozonation at pH7, 20 °C with 150% pick-up over different time from 0 to 60 min which can be observed in Fig. 3 as an example, its corresponding K/S, L^*, a^*, b^* values are listed in Table 1. It is clearly noted that each treated sample has an obvious difference from others in regard to color properties as treated by various ozonation processes, which on the other hand apparently indicated how complex the process parameters influence the color of dyed cotton fabric in ozonation. Table 2 exhibites the variation of the minimum, maximum, average and standard deviation of the dataset we used in the process modeling.

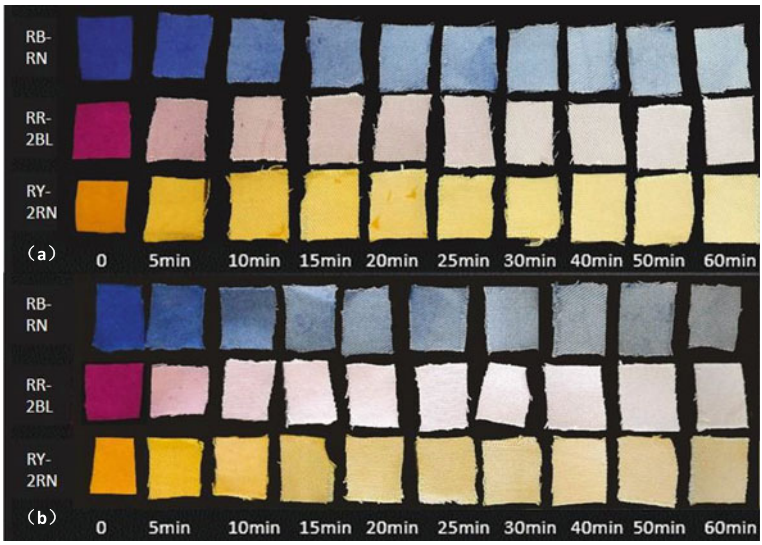


Fig. 3 The demonstration of the **a** front side and **b** back side of real ozone treated cotton samples

A total number of 612 sets of data collected in the experiment were divided into two groups, namely training data and testing data, where 75% of it was used for training and the remaining 25% was used for testing according to the general use of data division in the machine learning sector. As a result, 459 datasets (75%) were learnt by the models while the rest 153 datasets (25%) were distributed to test the model. The correlation of these factors (pH, Temperature, Pick-up and Time only as original color of fabric is not a continuous variable) to K/S, L^*, a^*, b^* values was estimated by Spearman rank correlation coefficients (based on (31)) and listed in Table 3. It is found that pH and treating time are slightly more relevant than temperature and pick-up in the ozonation that play a more important role.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{31}$$

where n is the number of observation, d_i is the difference between the ranks of corresponding variables. K-fold cross-validation (k = 10) was used in the modeling section. It is a popular statistical approach for estimating predictive models. Taking k = 10 as an example as it was the one we used in the modeling study, in which case 459 training sets of data would be divided randomly and equally into 10 disjoint folds, 9 folds of it would be split into training subset while the rest 1-fold would be used as validating subset. This procedure would be repeated 10 times with varied training and testing dataset at each time to validate the trained models. In order to evaluate the performance of models in validation, Mean Square Error (MSE) would be used based on:

Table 1 $K/S, L^*, A^*, B^*$ values of samples shown in Fig. 3

		Time (min)									
		0	5	10	15	20	25	30	40	50	60
RB-RN	K/S	8.12	3.70	2.17	1.63	1.24	1.17	0.89	0.85	0.66	8.12
	L^*	6.13	14.89	21.90	26.12	29.5	29.93	33.44	33.80	36.66	6.13
	a^*	−2.86	−4.47	−5.18	−5.68	−5.57	−5.27	−5.25	−5.19	−4.76	−2.86
	b^*	4.48	11.38	15.46	16.85	18.83	19.59	21.40	22.16	23.62	4.48
RR-2BL	K/S	1.10	0.95	0.70	0.66	0.52	0.57	0.51	0.39	0.39	1.16
	L^*	22.66	25.93	29.74	30.26	33.05	31.96	33.79	36.04	36.43	21.68
	a^*	−32.59	−37.32	−39.25	−40.16	−41.17	−41.82	−43.19	−42.44	−43.49	−32.39
	b^*	12.84	13.69	12.84	12.49	12.05	11.82	12.38	11.23	11.69	13.41
RY-2RN	K/S	6.60	3.70	2.90	2.20	1.81	1.58	1.16	1.11	0.87	6.60
	L^*	8.02	8.49	9.80	10.82	11.60	13.19	13.96	14.14	15.43	8.02
	a^*	20.41	15.35	12.99	11.47	9.59	8.62	7.31	6.52	5.56	20.41
	b^*	−17.99	−28.27	−32.09	−37.32	−40.95	−41.58	−47.94	−48.92	−52.39	−17.99

Table 2 The maximum, minimum, average and standard deviation of ozonation parameters

Parameters	Minimum	Maximum	Average	Std. dev.
Color[a]	0 (*Blue*)	1 (*Yellow*)	–	–
pH	1	13	7	3.463
Temperature (°C)	0	80	40	24.91
Pickup (%)	0	150	75	58.78
Time (min)	0	60	30	20.77
K/S	0.1	22.82	7.18	7.94
L^*	0.99	65.27	33.52	17.61
a^*	−58.99	53.68	−1.81	32.75
b^*	−90.53	88.04	3.88	42.34

[a]As color of fabric is not a continuous variable, it was represented in the bipolar form as 0 (blue), 0.5 (red) and 1 (yellow)

Table 3 The maximum, minimum, average and standard deviation of ozonation parameters

	pH	Temperature	Pickup	Time
K/S	0.0383	0.0648	−0.4862	−0.7913
L^*	−0.0579	0.1146	−0.1557	−0.3137
a^*	0.0248	0.1490	−0.4500	−0.7430
b^*	−0.0302	−0.0155	−0.0184	−0.0461

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i - p_i)^2 \tag{32}$$

where e_i is the real experimental results, whereas p_i is the predicted output of the specific model. Additionally, four statistical performance criteria, including MAE, RMSE, R and MRAE are used in this study for indicating the predictive performance of the obtained models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i - p_i| \tag{33}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - p_i)^2} \tag{34}$$

$$R(e,p) = \frac{\sum_{i=1}^{n}(e_i-\bar{e})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{n}(e_i-\bar{e})^2 \cdot \sum_{i=1}^{n}(p_i-\bar{p})^2}} \tag{35}$$

$$MRAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i-p_i|}{e_i} \tag{36}$$

The models' development and constructing were carried out using MATLAB R2015b for multi-output ELM and MLS-SVR, but R studio for MRF respectively on a laptop (Core i7-4710, 2.5 GHz, 16 GB RAM). All of the original data was regularized to the range of [0, 1] before using.

4 Results and Discussion

4.1 Modeling Training

4.1.1 ELM Models

ELM models with hidden nodes from 1 to 200 activated by Sigmoid, Sine and Hardlim functions are investigated respectively (the corresponding validation MSE is illustrated in Fig. 4 with a detailed demonstration of the trained ELM models possessing nodes from 1 to 140 in detail). The overfitting situation of ELM activated by Sigmoid and Sine is easy to be observed that starts from the ones with nodes around 100. More specifically, it is noted that Sigmoid trained ELM models performed similarly to the ones trained by Sine since MSE of these models both dropped as well as minimized at the ones with around 50 nodes (MSE $\approx$ 0.052) following by a dramatic enhancement. By contrast, validation MSE of Hardlim activated models performed generally stable with the growing number of nodes in the ELM model, but strictly a minimum of MSE $\approx$ 0.069 (lager than Sigmoid and Sine) at the one with 97 nodes still can be discovered in Fig. 4. Similar comparative results of the use of these activation functions in ELM can be found as well in the work of Singh and Balasundaram [28].

The use of activation functions in an artificial neural network is to convert an input signal of the node to an output signal by mapping non-linear properties. It is very important for an ELM model to learn and achieve the complicated mapping of input and output data by activating the nodes with a certain activation function. The graph of the activation functions we used is given in Fig. 5. It is noted that Sigmoid and Sine have much in common with their S-shaped curve and both are infinitely differentiable functions which make them easy to be understood and applied. However, on the other

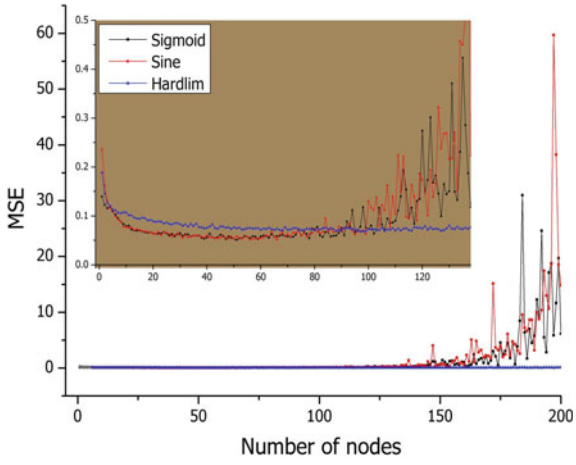


Fig. 4 Validation MSE of ELM models activated by different functions

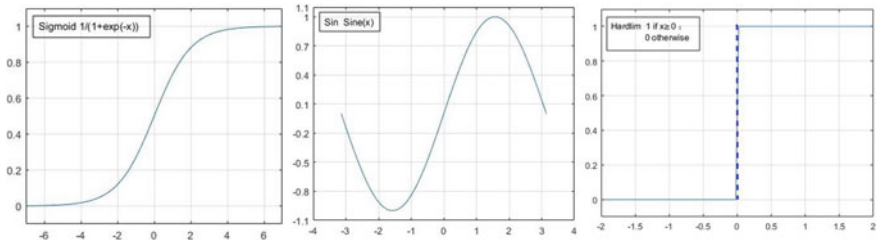


Fig. 5 Activation functions of ELM

hand, it may also result in their similar proximity and disadvantage in the ELM models as we can see their similar performance variation and the overfitting situation with the increasing nodes in Fig. 4. Hardlim performed least compared with Sigmoid and Sin in terms of their activated ELM models in this issue probably is owing to its oversaturation.

4.1.2 SVR Models

Multi-output SVR models with kernel functions of Linear, Sigmoid, Polynomial, RBF and ERBF were trained and developed using MLS-SVR toolbox. The corresponding results of minimum validation MSE are 0.05678, 0.00932, 0.08613, 0.00493 and 0.0092 respectively (as demonstrated in Fig. 6). It is worth noting that models trained with Linear kernel and Polynomial kernel are found that performed far poorly than the others. Performance of the ones with Sigmoid kernel and ERBF kernel are very close in a quite low level though the validation MSE of them is nearly two times than the SVR model with RBF kernel (which performed utmost in the comparison in this issue when its parameters are optimized to $\gamma = 32{,}768$, $\lambda = 9.7656e^{-4}$ and $p = 0.125$. For more information regarding the LOO optimization process used in the toolbox for these kernel parameters see [25]).

The kernel function is to transform the data as input into the required form to facilitate a non-linear decision surface to be a linear equation in higher dimensions where the computational power of the learning machine is heightened. The type of kernel function used would influence many of the characteristics of the SVR model. A wide range of kernels exist and it is hard to explain their individual characteristics, but it is well known that RBF kernel is recommended to be tried first in an SVR model due to the fact that it not only possesses certain similar parameters and behaviors of Linear and Sigmoid but also has fewer hyper parameters than Polynomial to complex the model. RBF is assumed as having computational advantages over other kernels depending on its easier and faster to compute the kernel values [29]. The lowest MSE it achieved in this case validates its preferential suitability to be employed in this study, and it should be attributed to that we have not too many features in the model but with comparatively large numbers of observations.

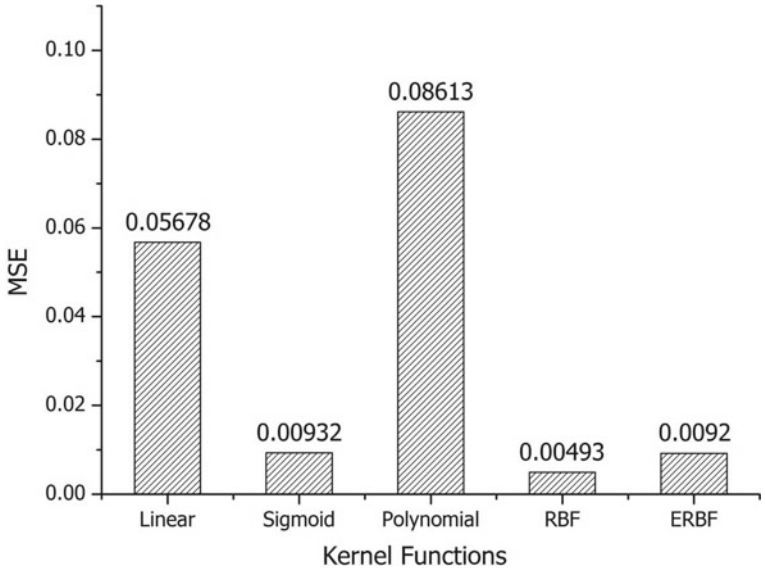


Fig. 6 Validation MSE of SVM models with varied kernel functions

RF models with different *mtry* (from 1 to 5), *minleaf* (from 1 to 10) and *ntree* (from 1 to 100) are trained and developed respectively, and the validation MSE of these models are given in Fig. 7 with a detailed demonstration of the ones $mtry = 1$ and *ntree* ranging from 1 to 100 excluding those which validation MSE higher than 0.026. In Fig. 7, the number of *mtry* in each regression tree node is found that plays a very significant role in affecting the models' prediction accuracy of the color properties of ozone treated cotton fabrics.

The falling curves of MSE with growing number of *mtry* may reveal that the five inputs we used to construct these RF modes, i.e. (1) color of dyed cotton, (2) pH, (3) temperature, (4) pick-up, (5) treating time of ozonation process, have a very clear independent relation with each other. As a result, RF models with five randomly selected features generally lead the low validation MSE in this comparison. It is also found that *ntree* played another significant role in RF models as MSE of these models decreased dramatically when the number of trees increased in the forest from 1 to 30. In general, these models perform steadily when there are more than 30 regression trees in the forest construction no matter what are the *mtry* or *minleaf* employed, but in order to save time and cost less in the model training process, 10 trees forest is sufficient and may be more recommended to be used in the color fading ozonation of dyed textile predicting model for further experiments. However, unlike the *mtry* and *ntree*, minimum number of samples in the leaf node, i.e. *minleaf* seems to be preferable to be less though it is relatively uninfluential. Depending on the observation of the detailed-depicted MSE plots of $1 - mtry$ RF models in Fig. 7, we

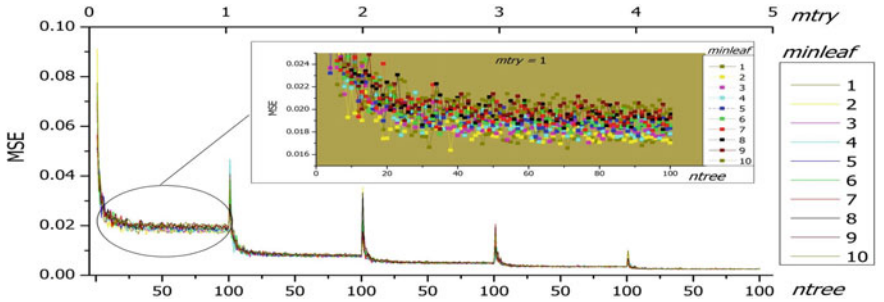


Fig. 7 MSE of RF models with varied number of features, leaves and trees

can see that the average MSE of achieved RF models generally enhanced when the number of leaves increased from 1 to 10.

4.2 Prediction Performance

The quality of a model is not only determined by its ability to learn from the data but also its ability to predict unseen data. Which two are so called learning capacity and generalization ability of a model. Models are seldom good in both of these two capacities. According to the training results obtained above, we can find out that Sine and Sigmoid trained ELM models have very similar performance that both optimized at 50 nodes, while SVR with RBF kernel function and RF as $mtry = 5$, $ntree = 10$, $minleaf = 2$, by contrast, clearly precede to all the others in their training process respectively. In order to further comparatively investigate the potential application of these three techniques without losing significant observations, the two ELM models were taken into account together with the RBF-SVR and the optimized RF ($mtry = 5$, $ntree = 10$, $minleaf = 2$) in this section.

To estimate and compare these optimized models, the prediction test using the testing dataset (which has not been used in the training and validation processes) is carried out. Table 4 presents a comparison of the prediction performance of ELM, SVR and RF models. It is found that, in general, ELM models using activation functions of Sigmoid (MSE = 0.0172) and Sine (MSE = 0.0173) do not make any big difference in regard to their prediction performance, but both of it are slightly poorer comparing with SVR and RF models. However, ELM models are the fastest-trained ones in the comparison, which means ELM model is still worth to be applied in certain resource-limited cases especially while limited training time is concerned. The most accurately-predicted model we can see, according to the finding in Table 3, is RF as it leads to the least testing error with higher R (0.9847) and less MSE (0.0036), MAE (0.0295), RMSE (0.0601) and MRAE (0.0062). However, it is also noted that training RF model requires a much longer time than the others (21s). As a result, it is worth taking the SVR models into account as it achieved the second

Table 4 Prediction performance of optimized models

Parameters	ELM-Sigmoid	ELM-Sine	SVR	RF
Training time (s)	0.0312	0.0936	0.9360	21
Average error (%)	0.1527	0.1549	0.1530	1.149
Maximum error (%)	0.3752	0.3700	0.3503	0.302
Minimum error (%)	0.0300	0.0029	0.0304	0.038
MSE	0.0172	0.0173	0.0043	0.0036
MAE	0.0894	0.0921	0.0429	0.0295
RMSE	0.1311	0.1315	0.0656	0.0601
R	0.9052	0.9063	0.9777	0.9847
MRAE	0.0197	0.0168	0.0109	0.0062

lowest error (R = 0.9777, MSE = 0.0043, MAE = 0.0429, RMSE = 0.656, MRAE = 0.0109) with a more acceptable shorter training time (0.9360 s).

Table 4 demonstrates the overall performance of the constructed models in terms of certain estimation evaluation indexes, but the detail of these predictions is neglected. As known that the constructed models possess four outputs, i.e. K/S, L^*, a^*, b^* values of reactive-dyed cotton fabrics treated in the color fading ozonation. How these predictive models work in detail with them is unclear. In order to reflect the real prediction performance (using testing data) of each trained model on predicting each single output separately, the predicted results range from output1 (k/s value) to output4 (b^*) versing real experimental data (target 1–4) is illustrated in Fig. 8a, b, c, d respectively.

In Fig. 8, the predicted values of models generally agree with the actual values, though the predictive errors varied in different levels for different models. As we can see that the gap of models' prediction performance is not that significant in Table 5 (taking MSE as an example, Sigmoid activated ELM = 0.0172, Sine activated ELM = 0.0173, SVR = 0.0043, RF = 0.0036), while the distribution of errors in terms of each single output prediction is observed that has a larger gap in the real application. In Fig. 8a, c, certain predicted data of ELM models can be clearly seen that is far different from the real target data in certain range, which situation would result in a big mistake in certain prediction application where good overall performance of the average of multiple outputs may hinder the discovery of a wrong prediction on specific single output. According to the linear fitting correlation coefficients of predicting data versing real experimental data (demonstrated in Fig. 6) listed in Table 5, the testing result obtained reveals that SVR ($R^2 = 0.9505$) model and RF model ($R^2 = 0.9555$) are actually more stable and suitable than ELM models ($R^2 = 0.8025$ and 0.8007 for Sigmoid and Sine activated respectively) in modeling color fading ozonation of dyed textile, in terms of overall prediction performance, and more importantly predicting multiple outputs without deviation on certain single output. This may

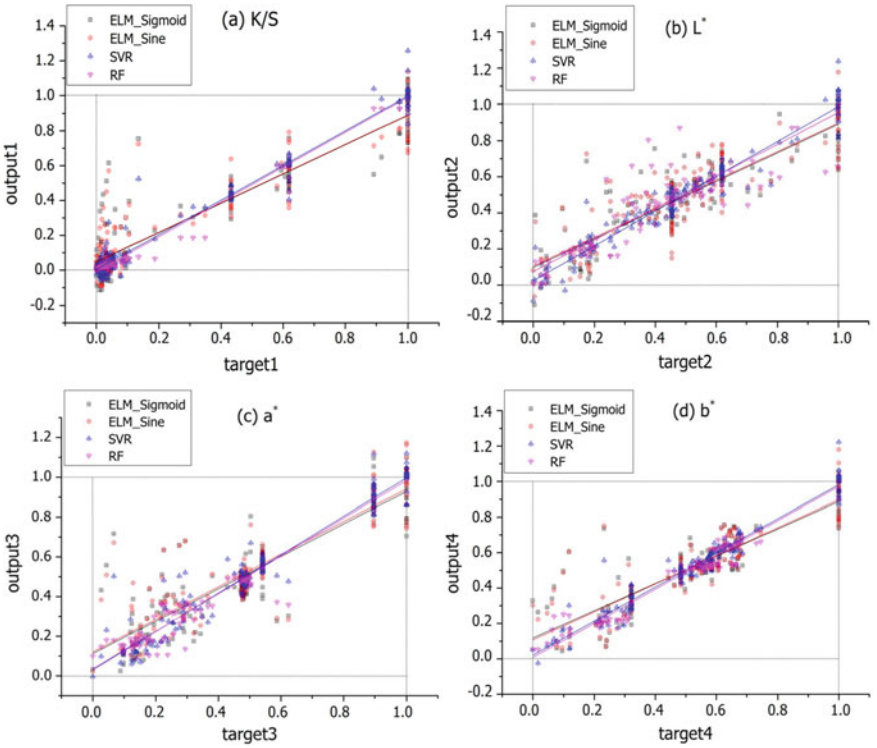


Fig. 8 Predicted data outputted by ELM (trained by Sigmoid and Sine respectively), SVR and RF versus experimental data

Table 5 Correlation coefficience of data in Fig. 8

R^2	ELM-Sigmoid	ELM-Sine	SVR	RF
Target 1-K/S	0.8474	0.8596	0.9683	0.9954
Target 2-L^*	0.7944	0.7517	0.9442	0.8816
Target 3-a^*	0.7903	0.8048	0.9380	0.9719
Target 4-b^*	0.7778	0.7868	0.9513	0.9731
Average	0.8025	0.8007	0.9505	0.9555

attribute to the features of data we used concerning color fading ozonation of dyed textile. While on the other hand, it could also attribute to a disadvantage of ELM that it completely relies on increasing the number of nodes to promote the predicting performance, which makes it risky to be applied in a complicated issue such as present investigation. The results also reveal that both SVR and RF can well deal with the interaction of variables and are comparatively more stable in multi-variable nonlinear modeling.

5 Conclusion and Prospective

In this chapter, three artificial technique modeling techniques i.e. ELM, SVR and RF were used to model the ozonation process for predicting the color properties of ozonated reactive-dyed cottons. The potential applicability of these models in the use of this textile finishing process modeling was estimated.

Color fading of dyed textile is a very vital process in the textile industry to obtain certain stylish effects on the product, which process has been increasingly used in last decades. Ozonation is a novel technology developed in recent years to be employed to achieve the color fading effect of textile with high performance not only in the respect of efficiency and quality but also in regards to the environmental sustainability. For the purpose of getting better understanding and application of color fading ozonation of textile in industrial scale, the complexity and nonlinearity of the factors and impacts of color fading ozonation on reactive-dyed cotton were investigated by process modeling. The effects of ozonation in terms of pH, temperature, water pick-up, treating time of process and dyed colors of fabrics on the color fading performance in terms of K/S, L^*, a^*, b^* values of reactive-dyed cotton were modeled using ELM, SVR and RF respectively. The finding results denoted that both of SVR and RF are potential applicable candidates for modeling the color fading ozonation process of dyed textile, as the predicted results of it on the ozonation process had a good agreement with the actual data entirely as well as individually. But taking the training time and cost as a consideration, SVR model would be more recommended than RF to be applied in real use. By contrast ELM models performed poorer in the prediction and were very unstable in terms of predicting certain individual output in multi-variable process modeling.

References

1. Kan CW, Cheung HF, Chan Q (2016) A study of plasma-induced ozone treatment on the colour fading of dyed cotton. J Clean Prod 112:3514–3524
2. Hoigné J (1988) The chemistry of ozone in water. In: Process technologies for water treatment. Springer, pp 121–141
3. He Z, Li M, Zuo D, Yi C (2019) Color fading of reactive-dyed cotton using UV-assisted ozonation. Ozone Sci Eng 41(1):60–68
4. He Z, Li M, Zuo D, Yi C (2018) The effect of denim color fading ozonation on yarns. Ozone Sci Eng 40(5)
5. He Z, Li M, Zuo D, Xu J, Yi C (2019) Effects of color fading ozonation on the color yield of reactive-dyed cotton. Dye Pigment 164
6. Suzuki K (ed) (2011) Artificial neural networks—industrial and control engineering application
7. Wu Y, Zhang Y, Liu X, Cai Z, Cai Y (2018) A multiobjective optimization-based sparse extreme learning machine algorithm. Neurocomputing 317:88–100
8. Sun Z, Choi T, Au K, Yu Y (2008) Sales forecasting using extreme learning machine with applications in fashion retailing. Decis Support Syst 46(1):413–421
9. Vapnik V (2013) The nature of statistical learning theory. Springer Science and Business Media
10. Basak D, Pal S, Patranabis DC (2007) Support vector regression. Neural Inf Process Rev 11(10):203–224

11. Ghosh A, Chatterjee P (2010) Prediction of cotton yarn properties using support vector machine. Fibers Polym 11(1):84–88
12. Nurwaha D, Wang X (2011) Prediction of rotor spun yarn strength using support vector machines method. Fibers Polym 12(4):546–549
13. Güneşoğlu S, Yüceer M (2018) A modeling study of micro-cracking processes of polyurethane coated cotton fabrics. Text Res J 88(24):2766–2781
14. Yap PH, Wang X, Wang L, Ong KL (2010) Prediction of wool knitwear pilling propensity using support vector machines. Text Res J 80(1):77–83
15. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(December):18–22
16. Breiman L (2001) Random forests Leo, pp 1–33
17. Kwon BK, Won JS, Kang DJ (2015) Fast defect detection for various types of surfaces using random forest with VOV features. Int J Precis Eng Manuf 16(5):965–970
18. Venkatraman V, Alsberg BK (2015) A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes. Dye Pigment 114(C):69–77
19. Kubelka P (1931) Ein Beitrag zur Optik der Farbanstriche (Contribution to the optic of paint). Zeitschrift fur Tech Phys 12:593–601
20. He Z, Tran KP, Thomassey S, Zeng X, Xu J, Yi C (2019) Modeling color fading ozonation of reactive-dyed cotton using extreme learning machine, support vector regression and random forest. Text Res J
21. Huang G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501
22. Rao CR (1971) Generalized inverse of matrices and its applications, no 04; QA263, R3
23. Mu K, Smola AJ, Scho B (1998) The connection between regularization operators and support vector kernels. Neural Netw 11:637–649
24. Scholkopf B, Smola AJ (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press
25. Xu S, An X, Qiao X, Zhu L, Li L (2013) Multi-output least-squares support vector regression machines. Pattern Recognit Lett 34(9):1078–1084
26. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22
27. Breiman L (2017) Classification and regression trees. Routledge
28. Singh R, Balasundaram S (2007) Application of extreme learning machine method for time series analysis. Int J Intell Technol 2(4):256–262
29. Smits GF, Jordaan EM (2002) Improved SVM regression using mixtures of kernels. In: Proceedings of the 2002 international joint conference on neural networks, IJCNN'02 (Cat. No. 02CH37290), vol 3. IEEE

Zhenglei He is currently a Ph.D. student studying in the laboratory of GEXTEX in ENSAIT. He received the M.S. degree in textile science and engineering from Wuhan Textile University. His research topic is focused on the modelling, simulation and optimization of textile manufacturing process via intelligent techniques.

Kim Phuc Tran obtained his Ph.D. degree in Automation and Applied Informatics at the Université de Nantes, Nantes, France, in 2016. He is currently an Associate Professor in Automation and Industrial Informatics at ENSAIT and GEMTEX, Roubaix, France. His research is focused on real-time anomaly detection techniques for Big Data, data mining, and machine learning techniques.

Sébastien Thomassey gained his Ph.D. degree in automation and information technology from the Lille I University, France, in 2002. He is currently associate professor (HDR) at ENSAIT and the GEMTEX laboratory. His works are published in refereed journals such as applied soft computing, decision support systems, European journal of operational research, and intentional journal of production economics. He co-edited the book "Artificial Intelligence for Fashion Industry in the

Big Data Era". He is actively involved in French and European collaborative research projects in production and supply chain management, sales forecasting and clustering/classification of fashion products.

Xianyi Zeng received his Ph.D. degree from the Centre d'Automatique, Université des Sciences et Technologies de Lille, France, in 1992. He is currently a Full Professor of ENSAIT, Roubaix, France, and director of the GEMTEX National Laboratory. He is also an Associate Editor of *International Journal of Computational Intelligence System and Journal of Fiber Bioengineering and Informatics*, a Guest Editor of Special Issues for six international journals, and a Senior Member of IEEE. He has organized 12 international conferences and workshops since 1996. Since 2000, he has been the leader of three European projects, four national research projects funded by the French government, three bilateral research cooperation projects, and more than 20 industrial projects. His research interests include: (1) intelligent decision support systems for fashion and material design, (2) modeling of human perception and cognition on industrial products, (3) intelligent wearable systems.

Changhai Yi is currently a professor (T2) at Wuhan Textile University, Wuhan, China. He is also the director of the National Local joint Engineering Laboratory for Advanced Textile Processing and Clean Production. Wuhan, China. He received his Ph.D. degree in Polymer Chemistry and Physics from Sun Yat-sen University, Guangzhou, China, in 1997. His research interests include: (1) Fiber modification and textile functional process, (2) Cleaner production and sustainable development of denim, (3) Informatization and intelligent manufacturing of denim.

Developing Alert Level for Aircraft Components

Wai Yeung Man and Eric T. T. Wong

Abstract In the aircraft industry maintenance is considered to be one of the key contributors to business success of an air carrier. By this, efforts are made to achieve maximum aircraft utilization with a reliability level as high as possible and minimal operating costs. As a result, aiming to increase profitability, and monitoring the reliability of an aircraft, its components and systems is of great benefit to aircraft carriers. As per Hong Kong Civil Aviation Department (HKCAD) requirements, all registered aircraft must have an approved maintenance schedule (AMS) to ensure aviation safety. In the AMS, system or component reliability plays an important role in condition-monitored maintenance program. To assist in the assessment of reliability of aircraft components, alert levels are established for the components which are to be controlled by the program. An alert level can help the operator to monitor engineering performance of an aircraft system or component during routine operations. Besides a consideration of the quality management processes to be contained within the AMS, this paper illustrates the development of an alert level for a helicopter air-conditioning system (ACS). The function of an ACS is to regulate the temperature, humidity, and air flow inside the helicopter. The reliability of the ACS is therefore important to an air operator. Once the alert level is triggered, approved maintenance actions need to be executed, such as component repair or replacement. In this paper, ACS of the helicopter McDonell Douglas 902 Explorer was chosen to illustrate the development of an alert level for the purpose of improving operational performance.

Keywords Helicopter air-conditioning system · Reliability · Alert level

W. Y. Man (✉) · E. T. T. Wong
Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China
e-mail: wai-yeung.man@connect.polyu.hk

E. T. T. Wong
e-mail: mmttwong@connect.polyu.hk

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_6

1 Introduction

In the early days of aviation, commercial airlines concerning with in-flight failures and component reliability developed maintenance schedules to help prevent costly occurrences. As aviation grew and advanced air transport technology arrived, safety, reliability and economics became important in order to attract passengers. It is obvious that to achieve a controlled balance between safety and economics, civil aviation regulatory authorities need to ensure minimum standards are maintained and a level playing field exists for fair competition among operators.

What to maintain, when to maintain and how to maintain, are the keys to the content of the maintenance schedules and a worldwide platform for airlines, aircraft manufacturers and regulatory authorities to share experience and knowledge on these particular issues in respect of new aircraft is being developed. Some commercial air transport operators may choose to maintain their aircraft in accordance with a maintenance program. This is a real-time system which consists of a maintenance schedule and a group of review and management procedures to ensure that it remains realistic and effective.

The Air Navigation (Hong Kong) Order 1995 [AN(HK)O] requires that an aircraft registered in Hong Kong in respect of which a certificate of airworthiness is in force, shall not fly unless it has been maintained in accordance with a maintenance schedule approved by the Hong Kong Civil Aviation Department (HKCAD) in relation to that aircraft. The term maintenance schedule [1] means the maintenance schedule together with any associated program, such as the reliability program, necessary to support the maintenance schedule. As the aircraft consists of an airframe, engine, propeller and other equipment, there are various sources of operational information. Not only will there be details of 'what and when' but also 'how' the parts are to be maintained; there are also details on the types of task that are to be performed. Such kind of details are published by the original equipment manufacturer (OEM) who may also be the type certificate holder (TCH) of that product. In aviation a type certificate is issued to signify the airworthiness of an aircraft manufacturing design or "type". The certificate is issued by a regulatory body, and once issued, the design cannot be changed. The certificate reflects a determination made by the regulatory body that the aircraft is manufactured according to an approved design, and that the design ensures compliance with national airworthiness requirements.

As data will be obtained from several manuals, there will be a collection of tasks to be accomplished at varying intervals. These intervals can be based either on flying hours, flight cycles or calendar time and sometimes there are combinations of these. It is quite often inconvenient to take each task as it comes and accomplishes it; it is usually expedient to parcel the tasks into packages of work that can be carried out when it is convenient to do so, but at a frequency not exceeding the approved intervals. The general rule that can be applied for compiling work packages is that tasks can quite often be done earlier than when recommended. They can only be done later with agreement of Hong Kong Civil Aviation Department and only in exceptional circumstances. So, for tasks that have more than one frequency in terms

of flying hours, flight cycles and calendar time, then the event that occurs first is normally the governing one.

The frequency of maintenance tasks is affected by the way the aircraft is to be operated. When the TCH recommendations are first compiled they would have in mind a 'typical' flight profile for the aircraft type; any deviation from this may need an adjustment on the basic recommendations. For example, an aircraft may have a 'typical' flight profile of three hours for every cycle while another may be of three cycles every hour. It can be seen that in these cases a schedule based solely on flying hours may mean the first aircraft is maintained too often and the other not enough, so, with the help of the TCH, usually a schedule can be developed for any particular type of operation.

The area of operation is another important consideration, for example operating over salt water may require special tasks, such as engine compressor washes and other maintenance, to be done on a more frequent basis. Similarly, operation in sandy areas or off rough strips may affect the tasks required. Likewise, the age of an aircraft may affect the number and frequency of tasks, particularly if it has ageing structural inspections and significant repairs.

Once established, an owner or operator may wish to amend the maintenance schedule due to addition or deletion of tasks or change of task interval. This can be done with the aid of an amendment to the maintenance schedule, which has to be submitted to HKCAD for agreement and subsequent approval. Doing tasks less frequently requires suitable justification in order that it may be approved. Proof that safety will not be compromised must be provided. Maintenance programs supported by a reliability program will have an advantage as they will readily be able to show how often a task has been performed without deterioration of the component or system.

Maintenance programs have long been based on the traditional method of maintaining safety margins by the prescription of fixed component lives and by aircraft 'strip-down' policies. The call for changes to the basic philosophy of aircraft maintenance has been greatly influenced by the present-day economic state of the industry as well as by changes in aircraft design philosophy allied to progress in engineering technology. These changes have, in turn, resulted in the necessity for the management and control of expensive engineering activities to take a new and more effective form. It is from this background that a maintenance process known as condition-monitored maintenance (CMM) has evolved. CMM [1] uses data on failures as items of 'condition' information which are evaluated to establish a necessity for the production or variation of hard time and on-condition requirements, or for other corrective actions to be prescribed. Failure rates and effects are analyzed to establish the need for corrective actions. Indeed, CMM can be used in its own right to identify the effects of deterioration, in order that steps may be taken to maintain the level of reliability inherent in the design of the item.

Although CMM accepts that failures will occur, it is necessary to be selective in its application. The acceptance of failures may be governed by the relative unimportance of the function, or by the fact that the function is safeguarded by system redundancy. To assist in the assessment of reliability, alert levels are established for the items which

are to be controlled by the maintenance program. When the alert level is exceeded an appropriate action has to be taken. It is important to realize that alert levels are not minimum acceptable airworthiness levels. Alert levels should, where possible, be based on the number of events which have occurred during a representative period of safe operation of the aircraft fleet. They should be updated periodically to reflect operating experience, product improvement, changes in procedures, etc.

When establishing alert levels based on operating experience, the normal period of operation taken is between two- and three-years dependent on fleet size and utilization. The alert levels will usually be so calculated as to be appropriate to events recorded in monthly or three-monthly periods of operation [2]. Large fleet will generate sufficient significant information much sooner than small fleets. Research on helicopter alert levels is scarce and it was not based on actual operating data. This is mainly because they are proprietary information.

2 Hong Kong Government Approved Maintenance Activities

HKCAD approves three primary maintenance activities.They are Hard Time, On-Condition and Condition Monitoring.

(i) Hard Time

This is a preventative process in which known deterioration of an item is limited to an acceptable level by the maintenance actions which are carried out at periods related to time in service (e.g. calendar time, number of cycles, number of landings). The prescribed actions normally include servicing and such other actions as overhaul, partial overhaul, replacement in accordance with instructions in the relevant manuals, so that the item concerned (e.g. system, component, portion of structure) is either replaced or restored to such a condition that it can be released for service for a further specified period.

For example, on the McDonnel Douglas helicopter model MD902, rotating swashplate assembly and non-rotating swashplate assembly are life limited to 7285 h and 1800 h respectively. Non-rotating swashplate is located directly above the mixer and below the rotating swashplate. The non-rotating swashplate is constructed of forged 7050 aluminum alloy. It receives collective and cyclic inputs and coordinates longitudinal and lateral flight control. It also transmits the control inputs to the rotating swashplate via the swashplate bearing.

Rotating swashplate is located directly above and assembled to the non-rotating swashplate by the swashplate bearing. The rotating swashplate transfers control inputs to the pitch links. It is in five-star configuration and made of forged 7050 aluminum alloy.

Rotating swashplate keeps rotating at 392 rpm once engines start, driven by a main rotor driveshaft. Non-rotating swashplate sliding up and down, tilting left and right frequently during flight. They are the most critical controls of the helicopter. When

the prescribed time reaches, the mentioned parts have to be replaced. Otherwise, the helicopter is not airworthy (Fig. 1).

(ii) On-Condition

This also is a preventative process but one in which the item is inspected or tested, at specified periods, to an appropriate standard in order to determine whether it can continue in service (such an inspection or test may reveal a need for servicing actions). The fundamental purpose of on-condition is to remove an item before its failures in services. It is not a philosophy of "fit until failure" or "fix and forget it".

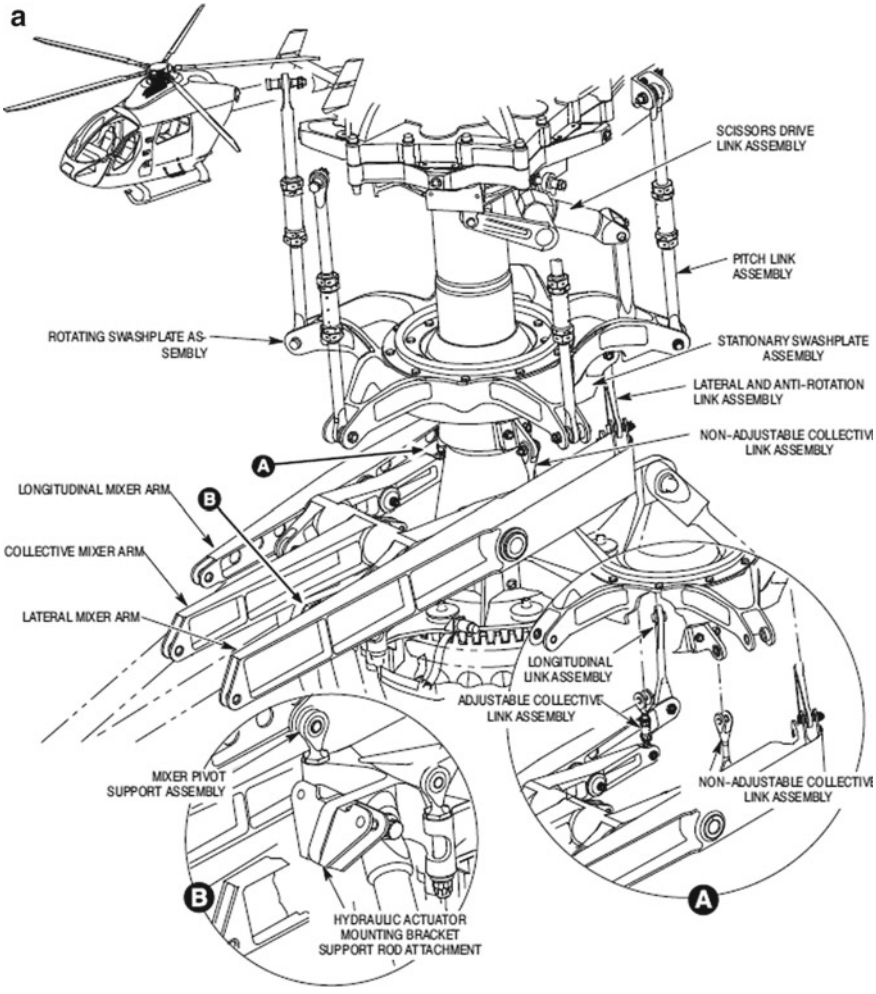
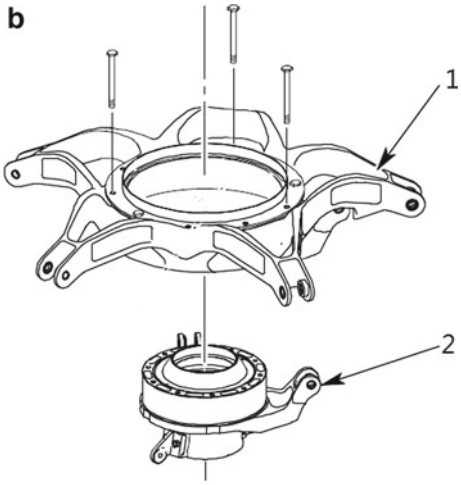


Fig. 1 **a** Rotating and non-rotating swashplate assemble. **b** Rotating and non-rotating swashplate

Fig. 1 (continued)



For example, on helicopter model MD902, flexbeam assembly and main rotor centering bearing assembly are on-condition items. They are inspected every 50 and 100 h. If damages are found on the mentioned items, they may be repaired or replaced depending on the nature of the damages. Otherwise, the helicopter is not airworthy. For the flexbeam, if a crack is found more than 2 inches, it needs to be replaced. For centering bearing, if cracks, rubber particles, housing not aligning with center bushing or separation between elastomeric material and the housing are found, it needs to be replaced.

Flexbeam is attached between the rotor blade and the main rotor hub. Flexing of the beam allows rotor blades to move about the feathering, flapping, and lead/lag axes without the need for hinges. The flexbeam is a composite structure made of laminated fiberglass. Flexbeam attachment points are hardened stainless steel bushings embedded in the laminate. An elastomeric flexbeam bumper is bonded around the flexbeam to reduce flexbeam defections during rotor starts and stops.

Centering bearing is installed at the hub end of the pitchcase. The centering bearing fits onto a snubber bolt in the lower hub half and provides a fixed feathering axis for the rotor blade. The centering bearing is captured between an upper and lower elastomeric damper assembly that help the flexbeam dampen out the lead/lag motion of the rotor blade (Fig. 2).

(iii) Condition Monitoring

This is not a preventative process, having neither Hard Time nor On-Condition elements, but one in which information on item gained from operational experience is collected, analysed and interpreted on a continuing basis as a means of implementing corrective procedures. Condition monitoring is repetitive and continuous, the key

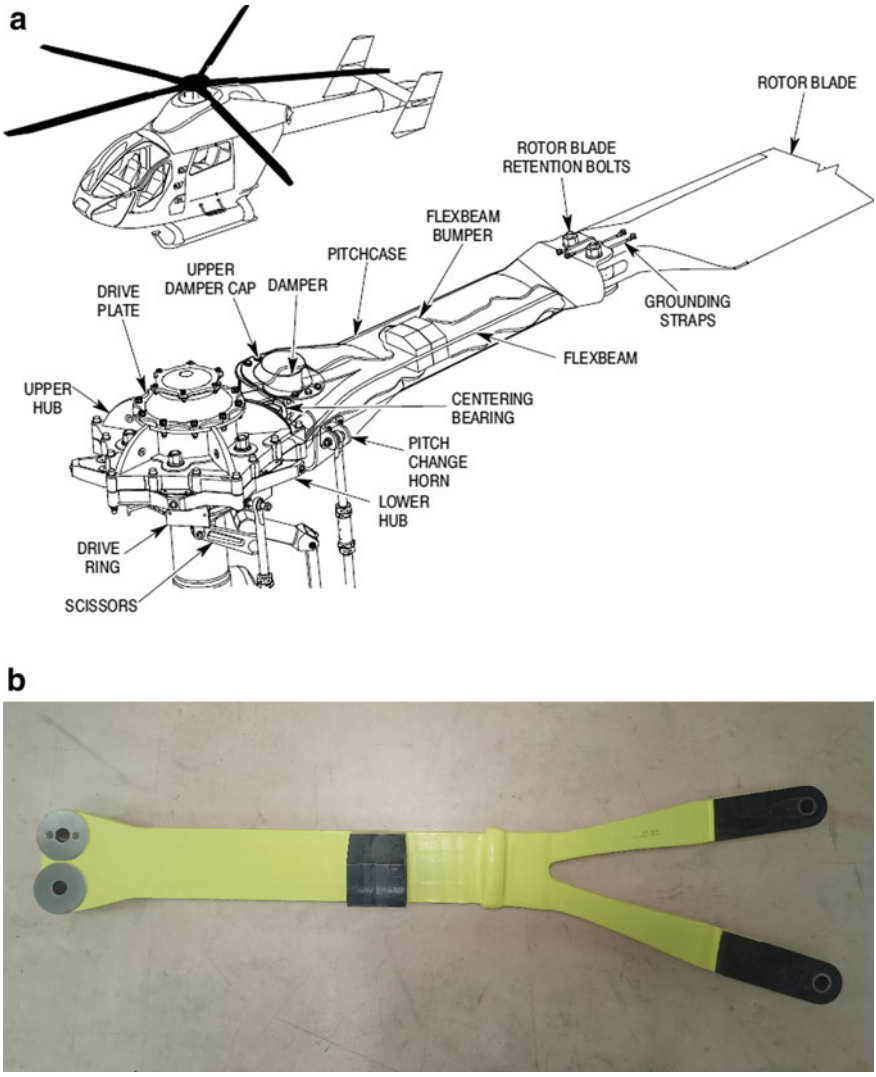


Fig. 2 **a** Flexbeam and centering bearing illustration. **b** Brand new flexbeam assembly. **c** Cracked flexbeam—out of tolerance (more than 2 in.)—unserviceable. **d** Cracked flexbeam—within tolerence (less than 2 in.)—serviceable—keep monitoring (a little black dot was marked to indicate the propagation of the crack). **e** Centering bearing

factor in its use being the introduction of aircraft embodying failure tolerant designs, which allow for replacement of some traditional failure preventative maintenance techniques by non-preventative techniques. Condition monitoring is not a relaxation of maintenance standards or of airworthiness control. In fact, it is more demanding

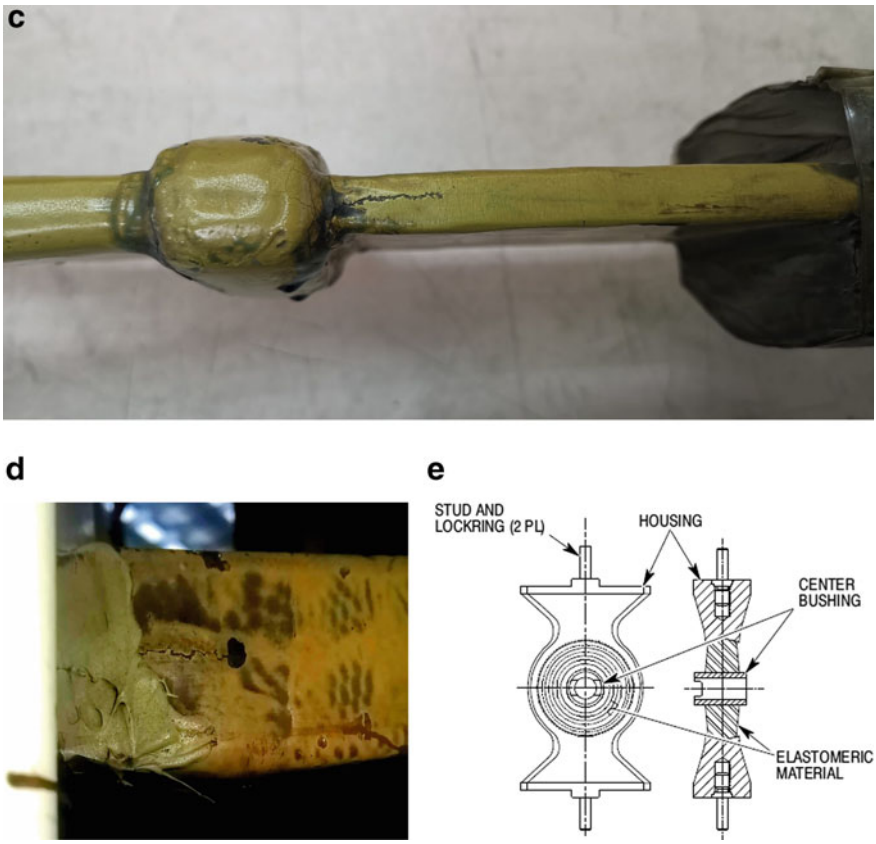


Fig. 2 (continued)

of both management and engineering capabilities than the traditional preventative maintenance approaches (Fig. 3).

3 Condition Monitored Maintenance Program

A maintenance program which provides for the application of Hard Time, On-Condition and Condition monitoring program is known as a Condition Monitored Maintenance Program (CMMP). There are two basic functions. Firstly, by means of the statistical reliability element, to provide a summary of aircraft fleet reliability and thus reflect the effectiveness of the way in which maintenance is being done by means of statistical reliability element. Secondly, to provide significant and timely technical information by which improvement of reliability may be achieved through changes to the program or to practices for implementing it. A properly managed

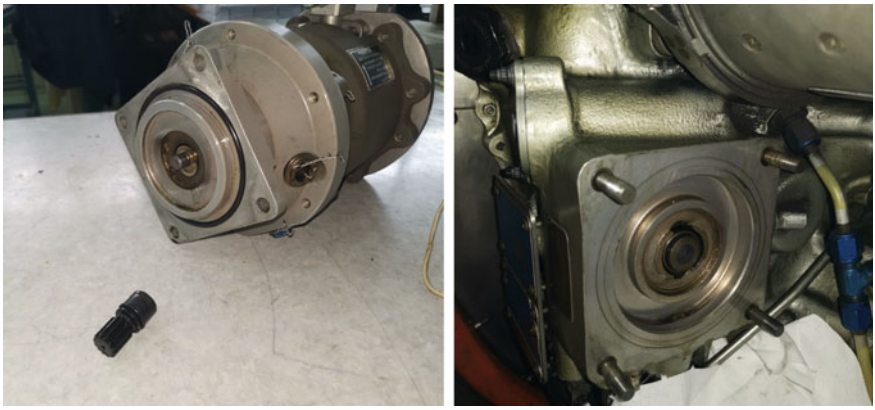


Fig. 3 Compressor split shaft separated from internal drive shaft

CMMP can contribute not only to the continuous airworthiness, but also improve the fleet reliability, produces better longterm maintenance planning, and of course, reduces the fleet operation and maintenance costs.

The key factors leading to a successful CMMP are the manner in which it is organized and the continuous monitoring of it by professionally qualified personnel. Within the CMMP, each reliability program has a controlling body, usually known as reliability control committee and it is responsible for the implementation, decision making, and day-to-day running of the program.

The reliability control committee would ensure that the reliability program establishes not only close co-operation between all relevant departments and personnel within the organization, but also liaison with other appropriate organizations. A typical communication and data flow chart is shown in Fig. 4:

The committee will have full authority to take actions to implement the objectives and processes defined in the CMMP. Normally, the quality manager or engineering manager will lead the committee and responsible to the Director of HKCAD for the program operation.

The committee should meet frequently to review the progress of the program and to discuss and resolve, if any, current problems. As the reliability program is approved by HKCAD, formal meetings with HKCAD will be held at agreed interval to assess the effectiveness of the program.

4 Data Collection

As mentioned at the beginning, aircrafts/components operation data are needed for preparing the reliability program. Due to the complexity and variety of the reliability program, there are different ways to collect data.

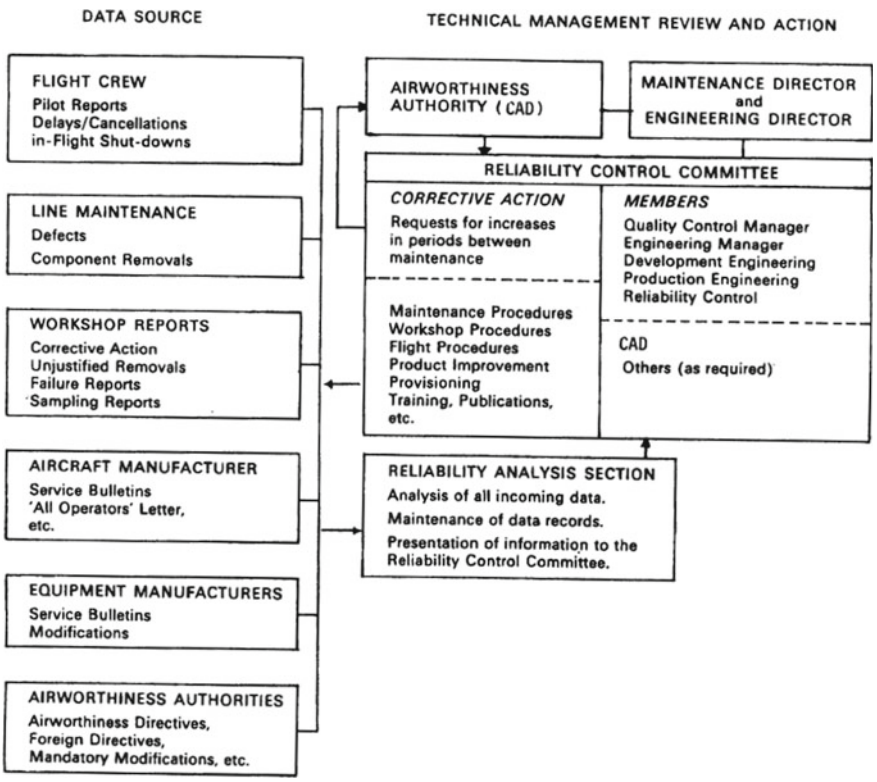


Fig. 4 A typical communication and data flow chart (*Source* CAD 418 [2])

(i) Pilot Reports

Usually known as "pireps", are reports of occurrences and malfunctions entered in the aircraft technical log by the flight crew for each flight. Pireps are one of the most significant sources of information, since they are a result of operational monitoring by the crew and are thus a direct indication of aircraft reliability as experienced by the flight crew.

(ii) Engine Unscheduled Shut-Downs

These are flight crew reports of engine shut-downs and usually include details of the indications and symptoms prior to shut- down. When analyzed, these reports provide an overall measure of propulsion system reliability, particularly when coupled with the investigations and records of engine unscheduled removals.

(iii) Mechanical Delays and Cancellations

These are normally daily reports, made by the line maintenance staff, of delays and cancellations resulting from mechanical defects. Normally each report gives the cause of delay and clearly identifies the system or component in which the defect

occurred. The details of any corrective action taken, and the period of the delay are also included.

(iv) Component Unscheduled Removals and Confirmed Failures

At the end of the prescribed reporting period the unscheduled removals and/or confirmed failure rates for each component are calculated to a base of 1000 or 500 h flying, or, where relevant, to some other base related to component running hours, cycles, landings, etc.

(v) Miscellaneous Reports

A variety of additional reports may be produced on a routine or non-routine basis. Such reports could range from formal minutes of reliability meetings to reports on the sample stripping of components, and also include special reports which have been requested by the maintenance management.

5 Alert Level Development

To assist in the assessment of reliability of aircraft components, alert levels are established for the components which are to be controlled by the program. Alert level is an indicator which when exceeded indicates that there has been an apparent deterioration in the normal behavior pattern of the item with which it is associated. When an alert level is exceeded appropriate maintenance actions should be taken as soon as possible.

When considering data based on components, it is useful to note that where a maintenance program is introduced to an aircraft fleet for the first time and in the early 'settling in' period, the number of failures which are not confirmed after an unscheduled removal can be as high as 40% for all components taken together. For individual components this can range from 5% for landing gear and flying control components to 65% for some communications and avionic components; thus, indicating the need for inclusion of data on both unscheduled removal and confirmed failure of components.

Typical acceptable procedures for establishing alert levels are outlined below. It can be seen that the resultant alert levels can vary according to the method of calculation, but this need not necessarily be considered to be of significance.

For each of the following sample calculations, a minimum of twelve months' operating data has to be available.

(i) Method 1

The three-monthly running average data for each system (or sub-system) is averaged over the sample operating period and is known as the Mean; the Mean is multiplied by 1.30 to produce the alert level for the given system. This is sometimes known as the '1.3 Mean' or '1.3 x' method.

(ii) Method 2

The Mean, as in Calculation 1, plus 3 Standard Deviations of the Mean.

(iii) Method 3

The mean, as in calculation 1, plus the Standard Deviation of the 'Mean of the Means', plus 3 Standard Deviations of the Mean.

6 Recalculation of Alert Level

If a significant change in the reliability of the particular item is experienced due to known actions such as modification, changes in maintenance or operating procedures, alert level of that item has to be recalculated. Any significant change in alert level requires written approval from HKCAD.

7 Corrective Actions

Corrective actions taken to improve the reliability of systems and components, and ultimately that of the fleet, will vary considerably and may typically include one or more of the following:

(i) Changes in operational procedures or improvements in fault finding techniques.
(ii) Changes to the scope and frequency of maintenance processes which may involve servicing and inspection, system tests or checks, overhaul, partial overhaul or bench testing or the introduction or variation of time limit, etc.
(iii) Modification action.
(iv) Non-routine inspections or adjustment.
(v) Change of materials, fuels and lubricants.
(vi) Use of different repair agencies.
(vii) Use of different sources of spares.
(viii) Variations of storage conditions.
(ix) Improvements in standards of staff training and technical literature.
(x) Amendment to the policy/procedures of the reliability program.

8 Essential Qualities of the CMMP

Condition monitored maintenance/reliability programs can vary from the very simple to the very complex, and thus it is impractical to describe their content in detail.

However, for HKCAD approval the CMMP has to be such that the considerations listed in (i)–(ix) are adequately covered:

(i) It generates a precise, specific and logical Quality assessment by the operator of the ability of the organization to achieve the stated objectives.
(ii) It enables the director, as the airworthiness authority, to accept initially, and, with subsequent continued monitoring, to have confidence in, the ability of the organization to such an extent that the director can renew certificates of airworthiness, approve changes to the maintenance schedules, etc., in accordance with evidence showing that the objectives of the program are being achieved.
(iii) It ensures that the operator provides himself with quality management of his organization.
(iv) It provides the operator with a basis for the discharges of his moral and legal obligations in respect of the operation of aircraft.
(v) It enables the Director to discharge its duties and legal obligations in respect of the maintenance aspects of airworthiness, and, where applicable, to delegate certain tasks to the operator.
(vi) The manner of presentation has to be acceptable to the Director.
(vii) With (i)–(vi) in mind, it states the objectives of the program as precisely as in possible, e.g. "maintenance of designated components by reliability management in place of routine overhaul", "condition monitoring as a primary maintenance process", etc.
(viii) The depth of description of the details of the program is such that:

   (a) The details can be understood by a technically qualified person.
   (b) Those factors which require formal CAD acceptance of any changes are clearly indicated.
   (c) All significant non-self-evident terms are defined.

(ix) In respect of individuals or departments within the organization:

   (a) The responsibility for the management of the program, and
   (b) The procedures for revision of the program, are clearly stated.

9 Assessment of CMMP Document

Following questions have been suggested by HK CAD to assist the air operator in making a preliminary assessment of the adequacy of the CMMP document:

(i) Is the document to be physically contained within the approved maintenance schedule? If it is to be a separate document, is it satisfactorily linked with, and identified within the approved maintenance schedule?

(ii) Are the objectives of the program clearly defined? e.g. “maintenance of designated items by reliability management in place of routine overhaul”, “confidence assessment of overhaul periods”, “airworthiness/economic quality management of maintenance”.
(iii) Does the approved maintenance schedule clearly state to which items the program is applicable?
(iv) Is there a glossary of terms associated with the program?
(v) What types of data are to be collected? How? By whom? When? How is this information to be sifted, grouped, transmitted and displayed?
(vi) What reports/displays are provided? By Whom? To Whom? When? How soon following data collection? How are delays in publishing controlled?
(vii) How is all information and data analyzed and interpreted to identify aircraft actual and potential condition? By Whom? When?
(viii) Is there provision within the organization for implementation of corrective actions and is this identified within the document? How are implementation time periods, effects and time for effect manifestation provided for?
(ix) Is there a requirement that the approved maintenance schedule be amended, and is the method of doing so included in the program, e.g. variation of time limitations, additional checks?
(x) Is there a requirement that maintenance manuals be amended and is the method of doing so included in the program, e.g. maintenance practices, tools and equipment, materials?
(xi) Is there a requirement that the operations manual/crew manual be amended, and is the method of doing so included in the program, e.g. crew drills, check lists, defect reporting?
(xii) What provision is made for corrective action follow-up and for checks on compliance with original intention, e.g. those which are not working out in practice, spares provisioning, timetables for the incorporation of modifications?
(xiii) Who is responsible for the management of the document?
(xiv) Is there a diagram of the relationship between the departments and groups concerned with the program and does it show the flow of condition monitoring data, its handling and the prescribed reaction to it?
(xv) Are all of the departments involved in the program included and are there any responsibilities not allocated?
(xvi) What quality management processes are contained within the program in respect of:

    (a) Responsibility for the document itself and the procedure for its amendment?
    (b) Monitoring of the performance of the program by statistical reliability and other methods?
    (c) Committee consideration of program implementation and monitoring of performance?

(d) Consideration of reports on incidents and accidents and other events which can affect airworthiness?
(e) Program management and discipline?

To illustrate the development of alert levels for an aircraft system and its application in aircraft maintenance, a case study based on the MD 902 Explorer helicopters is shown below.

10 Case Study

Heliservices (HK) Ltd, the sole provider of rotary-wing services in Hong Kong, offers a wide range of aviation services including aircraft charter for sightseeing and VIP services, filming and photography, lifting works, powerline maintenance and methodical inspection. Heliservices also provides aircraft management and maintenance services and unmanned aerial systems (UAS).

Heliservices operates a fleet of three MD 902 Explorer helicopters. Each helicopter has an air-conditioning system mainly consists of a compressor, evaporator, condenser, receiver dehydrator, thermal expansion valve and pressure sensors. The air-conditioning system (ACS) supplies ventilation, temperature, and humidity control to the helicopter. In this case study the ACS of the MD 902 are chosen to evaluate the operational improvement opportunities.

From past record, the ACS has a high failure rate. Without ACS, passengers and crews experience a discomfort journey. To provide enjoyable journeys, operation department often needs to reschedule the flights for engineering department to rectify any defects relate to ACS. Alert level was chosen to monitor the system or component performance so that an optimal time for maintenance can be agreed between operation and engineering department. It was found that most of the time, ACS defects were mainly due to compressor failures.

11 Helicopter Air-Conditioning System

The effect of aircraft cabin temperature and ventilation on crew performance has long been a subject of aeromedical study by flight physiologists and physicians alike. Air conditioning systems for helicopters have been around for a long time. Some systems in fact still operate with the old CFC-12 (R12) refrigerant, while the newer vapor-cycle systems are serviced with the more environmentally friendly HFC-134a (R134A). For the MD902 Explorer, air-conditioned air is routed through the fresh air vent system gaspar assemblies. The cockpit has four gaspar assemblies and the cabin area has six. The air conditioning system is a vapor-cycle type system that provides conditioned air to cool the cockpit and cabin areas.

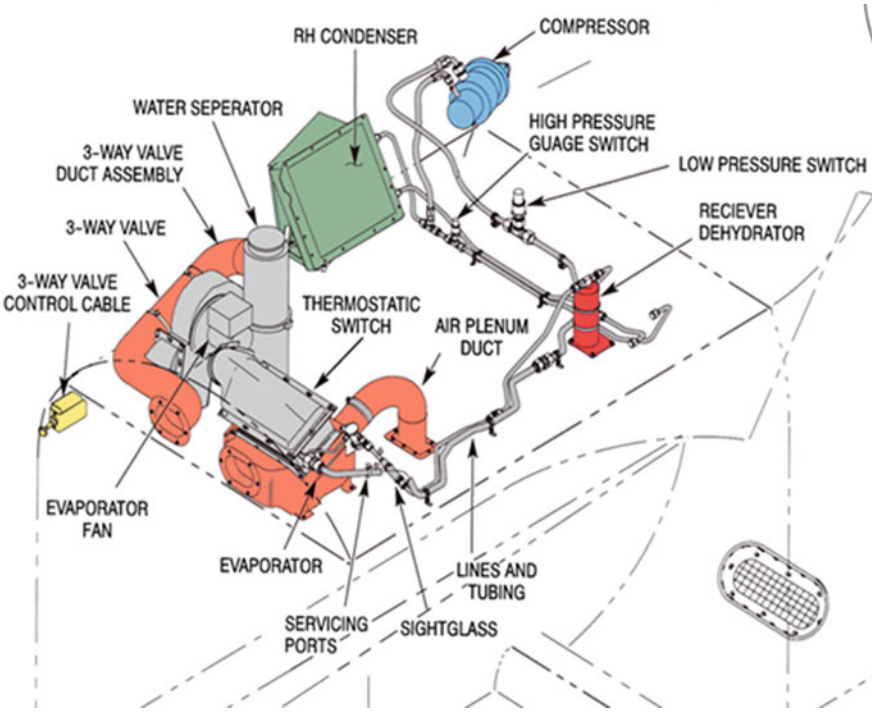


Fig. 5 MD902 air conditioning system layout

This section provides a general overview of a typical vapor-cycle air conditioning system and how it functions. This type of system operates in a closed loop, in which the refrigerant absorbs heat from the cabin and rejects it into the outside air. The refrigerant then returns to the cabin to repeat the cycle. The operation of the system is described below and involves these major components: compressor, condenser, receiver-dehydrator, and evaporator (Fig. 5).

The compressor receives through the suction port low pressure heat laden refrigerant vapor from the evaporator. The compressor pressurizes the refrigerant depending on system demand. This increases the temperature of the refrigerant. At this temperature and pressure the refrigerant is above its boiling point. The compressor then discharges it out of the discharge port.

The superheated refrigerant then flows into the condenser. In the condenser the high pressure vapor condenses into a high pressure liquid. This is achieved by using a fan to force air over the surface of the condenser enabling heat to transfer from the refrigerant to the outside air thus reducing its temperature. Ideally, only refrigerant in the form of a high pressure liquid leaves the condenser outlet.

The high pressure liquid refrigerant flows into the receiver-dehydrator which stores, dries and filters the liquid refrigerant. The liquid refrigerant then flows from the receiver-dryer to the expansion valves. The expansion valves change the refrigerant

into low pressure, low temperature liquid by lowering the pressure using a variable orifice. The orifice has high pressure on one side (from the receiver-dehydrator), and low pressure on the other side (evaporator and compressor), and allows a small quantity of refrigerant to flow through it. The sudden drop in pressure and temperature causes some of the refrigerant to vaporize. The low pressure, low temperature liquid then flows to the evaporator where the heat is transferred from its surface to the refrigerant through vaporization. The heat comes from inside the cabin and is blown over the evaporator's surface. Once the refrigerant has completely vaporized and reached its saturation point it should still be able to carry more heat.

The refrigerant continues to flow through the remainder of the evaporator coils absorbing more heat and becoming slightly superheated. Thus, air is blown through the evaporator, has its heat extracted/transferred to the refrigerant, resulting in cooled air being ducted into the cockpit and cabin through strategically placed air outlets.

The low pressure low temperature slightly superheated vapor refrigerant flows to the compressor and the cycle repeats itself.

The fresh air vent system provides outside airflow into the cockpit and cabin areas. Outside air enters the fresh air ducts and is available to the cockpit and cabin areas by manual adjustment of gaspar assemblies that provide flow volume up to 20 cubic feet per minute. The cockpit has fourgaspar assemblies and the cabin area has six. An integral two-speed vent fan can sustain fresh airflow at up to 180 cubic feet per minute [3]. Vent fan control is provided from the cockpit (Fig. 6).

12 Compressor Failures

As mentioned above, defects of the compressor have contributed to the failure of air conditioning system. The compressor internal drive shaft is connected to the waisted spline input shaft by a shear pin. If excessive torque is caused by the compressor or gearbox problem, the shear pin is sheared so that damage will not be transferred to the engine (Figs. 7, 8).

13 Alert Level Calculation

The number of air-conditioning compressor failures were collected among the fleet covered each 500 flying-hour period from January 2014 to December 2017. By using method 3 (The Mean plus the Standard Deviation of the 'Mean of the Means', plus 3 Standard Deviations of the Mean) as shown above, the alert level was calculated to be 12.7. This means that whenever there are 12.7 failures or more per 500 actual flying hours, appropriate maintenance action needs to be done. Detailed calculations are shown in Table 1:

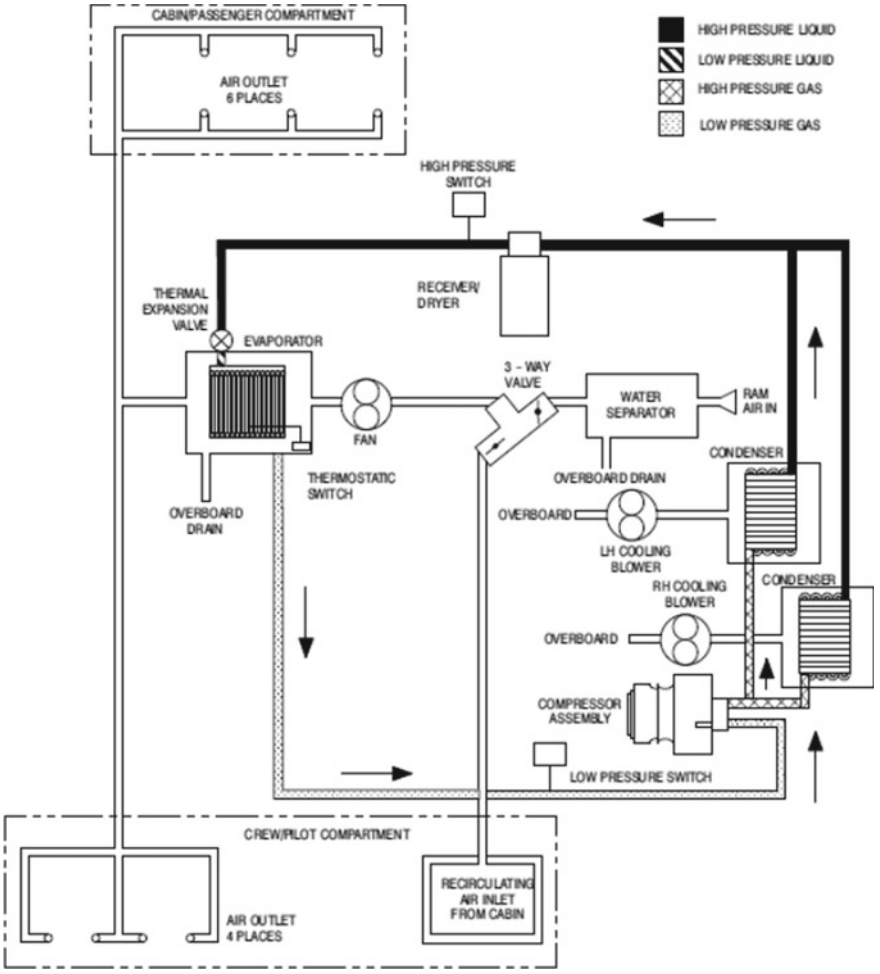


Fig. 6 MD902 air conditioning system operation [4]

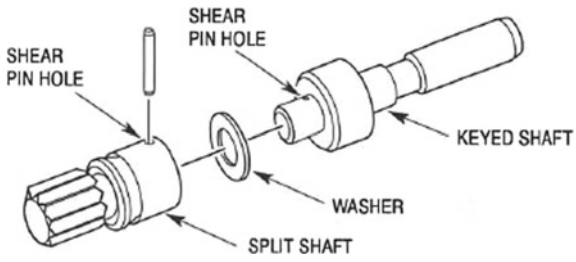


Fig. 7 Shear pin of air-conditioning compressor drive shaft

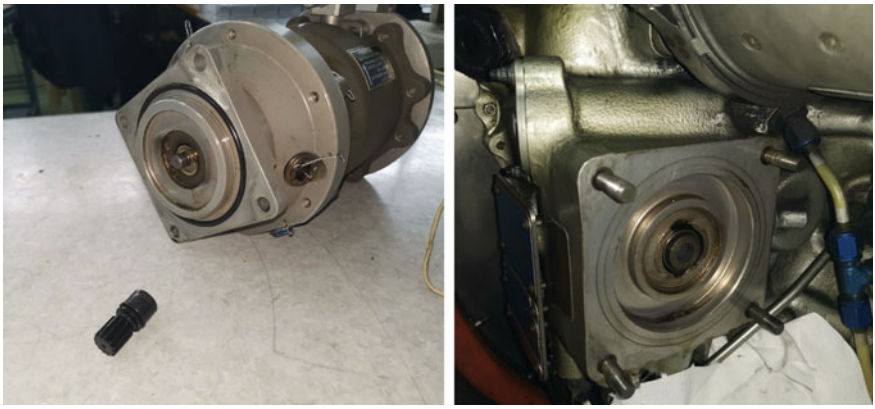


Fig. 8 Failure of compressor—shear pin broken; split shaft left in engine accessory gear

Table 1 2014–2017 MD 902 Explorer fleet air-conditioning compressor failures

Actual flying hours	Failures among fleet, (X)	$\lvert X - \overline{X}\rvert$	$\lvert X - \overline{X}\rvert^2$	Mean of X, (Y)	$\lvert Y - \overline{Y}\rvert$, D	$\lvert Y - \overline{Y}\rvert^2$, D^2
0–500	0	3	9			
500.01–1000	5	2	4	2.5	1.3	1.69
1000.01–1500	3	0	0	4	0.2	0.04
1500.01–2000	2	1	1	2.5	1.3	1.69
2000.01–2500	9	6	36	5.5	1.7	2.89
2500.01–3000	3	0	0	6	2.2	4.84
3000.01–3500	1	2	4	2	1.8	3.24

$N_1 = 7$

$\overline{X} = \frac{\sum X}{N_1} = 3.3$

$\sum \lvert X - \overline{X}\rvert^2 = 54$

$SD = \sqrt{\frac{\sum \lvert X - \overline{X}\rvert^2}{N_1 - 1}} = \sqrt{\frac{54}{6}} = 3$

3 X SD = 9

Mean of mean of $X = \frac{\sum Y}{N_2} = 22.5/6 = 3.75 \approx 3.8$ where $N_2 = 6$

$\sum D^2 = 14.39 \approx 14.4$

$\sum D = 8.3$

Standard deviation of Mean of Mean $= \sqrt{\frac{\sum D^2}{N_2 - 1} - \left(\frac{\sum D}{N_2 - 1}\right)^2} = \sqrt{\frac{14.4}{5} - \left(\frac{8.3}{5}\right)^2} = 0.4$

Alert level $= 0.4 + 3.3 + 9 = 12.7$.

Table 2 2014–2018 MD 902 Explorer air-conditioning compressor failures

Actual flying hours	Failures among fleet, (X)	$\lvert X - \overline{X} \rvert$	$\lvert X - \overline{X} \rvert^2$	Mean of X, (Y)	$\lvert Y - \overline{Y} \rvert^2$, D	$\lvert Y - \overline{Y} \rvert^2$, D^2
0–500.00	0	3	9			
500.01–1000	5	2	4	2.5	1.1	1.21
1000.01–1500	3	0	0	4	0.4	0.16
1500.01–2000	2	1	1	2.5	1.1	1.21
2000.01–2500	9	6	36	5.5	1.9	3.61
2500.01–3000	3	0	0	6	2.4	5.76
3000.01–3500	2	1	1	2.5	1.1	1.21
3500.01–4000	3	0	0	2.5	1.1	1.21

This alert level cautioned the engineering department and her staff started observing pilot procedure of operating the ACS during power assurance check. It was noted that pilots used to switch on air-conditioning system during engine power assurance check, which engaged the compressor immediately, and this required single engine operation at 75% torque. Eventually a new procedure came up and pilots were instructed not to engage air-conditioning system during engine power assurance check.

Based on the new procedure and additional operating data, the alert level 11.8 was achieved. This drop in alert level means that the performance of the air-conditioning compressor was improved. In other words, the compressor has become more reliable. Again, detailed calculations can be found in Table 2.

$N_1 = 8$

$\overline{X} = \frac{\sum X}{N_1} = 3.4$

$\sum \lvert X - \overline{X} \rvert^2 = 51$

$SD = \sqrt{\frac{\sum \lvert X - \overline{X} \rvert^2}{N_1 - 1}} = \sqrt{\frac{51}{7}} = 2.7$

$3 \text{ X SD} = 8.1$

Mean of Mean of $X = \frac{\sum Y}{N_2} = 25.5/7 = 3.64 \approx 3.6$ where $N_2 = 7$

$\sum D^2 = 14.37 \approx 14.4$

$\sum D = 9.1$

Standard deviation of Mean of Mean $= \sqrt{\frac{\sum D^2}{N_2 - 1} - \left(\frac{\sum D}{N_2 - 1}\right)^2} = \sqrt{\frac{14.4}{6} - \left(\frac{9.1}{6}\right)^2} = 0.3$

Alert level $= 0.3 + 3.4 + 8.1 = 11.8$.

14 Conclusion

Reliability engineering and maintenance are the source of aviation safety. They are one of the most important aims in the aviation industry. Modern reliability engineering is one way of reducing operating costs without compromising safety. Through Condition Monitored Maintenance, one can achieve better operational performance of aircraft systems at minimum costs.

In this chapter, an alert level was established based on the operating data of the chosen helicopter air-conditioning system compressor. Being cautioned by the high failure rate of the air-conditioning system, the maintenance engineering team proposed a new operating procedure for the helicopter pilots. For this new procedure, pilots were told not to engage air-conditioning system during engine power assurance check. Subsequently it was found that implementation of this procedure resulted in an improved utilization of the air-conditioning system, as evidenced through the new alert level - a reduction from 12.7 to 11.8. This case study illustrates that alert levels can be used to monitor and improve fleet operational performance effectively.

Notation	Description
D	Absolute value of the difference between mean of X and mean of mean of X
N_1	Number of ranges of actual flying hours
N_2	Number of values of mean of X
SD	Standard deviation
X	Number of failures among fleet
$\overline{X}$	Mean of number of failures among fleet
$\lvert X - \overline{X} \rvert$	Absolute value of the difference between number of failures among fleet and mean of number of failures among fleet
Y	Mean of X
$\overline{Y}$	Mean of mean of X

References

1. Hong Kong Civil Aviation Department (2014) CAD 452, aircraft maintenance schedules and programs. Government Logistics Department
2. Hong Kong Civil Aviation Department (2012) CAD 418, condition monitored maintenance: an explanatory handbook. Government Logistics Department
3. MD Helicopters (2014) Technical Description—MD 902 Explorer. Retrieved from https://www.mdhelicopters.com/files/Models/MD902_Tech_Desc.pdf. 14 Feb 2014
4. MD Explorer® (MDHI Model MD900®) Rotorcraft Maintenance Manual (CSP-900RMM-2) Chapter 21-50-00, 16 Sep 2019

Wai Yeung Man Simon received the Bachelor's degree of electronic engineering from The Chinese University of Hong Kong in 2011. He furthered his studies in mechanical engineering at The Hong Kong Polytechnic University since 2018. Simon is particularly interested in aviation maintenance and engineering. He is currently a licensed avionics engineer at Heliservices HK Ltd. He is specialized in helicopter avionics systems inspection, troubleshooting, fault rectification and system modifications. He also holds a helicopter airframe and engine license.

Eric T. T. Wong Eric received his M.Sc. degree in Plant Engineering from the Loughborough University(UK) and Ph.D. degree from the Leicester University (UK). Prior to joining the Mechanical Engineering Department of the Hong Kong Polytechnic University, he worked for the Hong Kong Civil Aviation Department in the area of aviation safety. His research interests include: aviation safety management, risk modeling, and quality management. He is a Fellow Member of the Royal Aeronautical Society (UK) and a Founding Member of the Aircraft Engineering Discipline, Hong Kong Institution of Engineers. He has been the Treasurer of the Royal Aeronautical Society (Hong Kong) since its inception in September 1983. He was Vice-Chairman of the HK Institute of Marine Technology in 2002 and Program Chair of the ISSAT International Conference on Modeling of Complex Systems and Environments (2007) at Ho Chi Minh City, Vietnam. He is currently an Adjunct Associate Professor of the Hong Kong Polytechnic University and an Associate Editor of the International Journal of Industrial Engineering—Theory, Application and Practice.

Computation in Network Reliability

Shin-Guang Chen

Abstract This chapter presents the primitive steps in computation of network reliability, and some of the popular examples by applying network reliability. Network reliability is one of the most interesting topics in network theory. It has been applied to many real-world applications such as traffic planning, computer network planning, power transmission planning, etc. The planners usually want to know how reliable the systems they planned. Network reliability helps them to get the answer of such questions. Due to its NP-hard nature, many approaches have been developed to tackle the efficiency problems in computations. One of the most promising methods is the three-stage approach, which divides the problem into three smaller problems and conquers them, making the resolution process more efficient than others. Several examples are presented in the chapter for illustrative purposes.

1 Introduction

Network reliability is one of the most interesting topics in network theory. It has been applied to many real-world applications such as traffic planning, computer network planning, power transmission planning, etc. The planners usually want to know how reliable the systems they planned. Network reliability helps them to get the answer of such questions. Since 1954, the maximal flow problems have gained much attention in the world [23]. They are also extended to many other fields for applications [14]. Aggarwal et al. firstly discussed the reliability problem without flow in a binary-state network [3]. Lee [17] extended it to cover the flow cases. Aggarwal et al. [2] presented the minimal path (MP) method to solve the network reliability in a binary-state flow network. Xue [26] began to discuss the reliability analysis in a multistate network (MSN), which is also referred to the stochastic-flow network (SFN). A SFN is a network whose flow has stochastic states. Lin et al. [18] illustrated the reliability calculation of a SFN in terms of MPs or minimal cuts (MC)s [15]. They also setup

S.-G. Chen (✉)
Tungnan University, New Taipei City, Taiwan
e-mail: bobchen@mail.tnu.edu.tw

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_7

the three-stages in these calculations: a. Searching for all MPs [10, 12, 24]/MCs [1, 7, 16, 30]; b. Searching for all $d-$MPs [20, 27]/$d-$MCs [28] from these MPs/MCs; c. Calculating union probability from these $d-$MPs [5, 6, 32]/$d-$MCs [12, 19]. Lin [20] greatly simplified the three-stages method (TSM) and developed more simple and efficient algorithm for the reliability evaluation of a general SFN.

Doulliez and Jamoulle [13] also proposed a novel method to evaluate network reliability, namely state space decomposition (SSD). However, their algorithm has some flaw, and cannot obtain the correct value. Aven [4] presented the correct one for state space decomposition approach, and till now still keeps excellent efficiency in SSD-based approach. The complexity of the worst cases for SSD is $O(N^d)$, where N is the number of arces, and d is the demand. TSM divided the reliability evaluation process into three stages. The first stage—finding MPs has the complexity $O(W + S)$ [10], where W is the number of MPs, and S is the number of cycles. The second stage—searching for d-MPs has the complexity $O(\prod_{k=1}^{u} r_k + |B| \binom{z+d}{d-1})$ [21], where r_k is the kth segment of u flow constraints, $|B|$ is the number of nodes, z is the number of MPs, and d is the demand. The third stage—calculating the union probability of d-MPs has the complexity $O(2^Q)$ [5, 6], where Q is the number of d-MPs. Therefore, when N is large, the evaluation efficiency of TSM is undoubtedly superior to SSD.

The remainder of the chapter is organized as follows. The mathematical preliminaries for the network reliability are presented in Sect. 2. The network representation method [9] is given in Sect. 3. The MP searching method [10] is given in Sect. 4. The d-MP searching method [21] is given in Sect. 5. The union probability calculation method [5] is given in Sect. 6. Finally, Sect. 7 draws the conclusion of this chapter.

2 Preliminaries

Let (A, B) be a SFN, where $A = \{a_i | 1 \le i \le N\}$ is the set of edges, B is the set of nodes. Let $M = (m_1, m_2, \ldots, m_N)$ be a vector with m_i (an integer) being the maximal capacity of a_i. The network is assumed to satisfy the following assumptions.

1. Flow in the network satisfies the flow-conservation law [14].
2. The nodes are perfect.
3. The capacity of a_i is an integer-valued random variable which takes values from the set $\{0, 1, 2, \ldots, m_i\}$ according to a given distribution μ_i.
4. The states in edges are statistically independent from each other.

2.1 Modeling of Networks

Assume that ρ_1, ρ_2, …, ρ_z are totally the MPs from the source to the sink. Thus, the network is modeled by two vectors: the state vector $X = (x_1, x_2, \ldots, x_N)$ and the flow vector $Y = (y_1, y_2, \ldots, y_z)$, where x_i denotes the current state of a_i and y_j

denotes the current flow on ρ_j. Then, Y is feasible if and only if

$$\sum_{j=1}^{z}\{y_j|a_i \in \rho_j\} \leq m_i, \text{ for } i = 1, 2, \ldots, N. \tag{1}$$

This constraint describes that the total flow through a_i can not exceed the maximal capacity of a_i.Then, the set of Y denoted $\kappa_M \equiv \{Y|Y \text{ is feasible under } M\}$.

Similarly, Y is feasible under $X = (x_1, x_2, \ldots, x_N)$ if and only if

$$\sum_{j=1}^{z}\{y_j|a_i \in \rho_j\} \leq x_i, \text{ for } i = 1, 2, \ldots, N. \tag{2}$$

For clarity, let $\kappa_X = \{Y|Y \text{ is feasible under } X\}$. The maximal flow under X is defined as $\varpi_X \equiv \max\{\sum_{j=1}^{z} y_j|Y \in \kappa_X\}$.

Then, the flow vector $Y \in \kappa_M$ could be found such that the total flow of Y equals d. It is defined in the following constraint,

$$\sum_{j=1}^{z} y_j = d. \tag{3}$$

Let $\mathbf{Y}$=$\{Y|Y \in \kappa_M$ and satisfies Eq. (3)$\}$. We have the following lemmas [20].

Lemma 1 *If X is a d-MP for d, then there is an $Y \in \mathbf{Y}$ such that*

$$x_i = \sum_{j=1}^{z}\{y_j|a_i \in \rho_j\}, \quad \text{for each } i = 1, 2, \ldots, N. \tag{4}$$

Given $Y \in \mathbf{Y}$, a state vector $X_Y = (x_1, x_2, \ldots, x_N)$ can be built by Eq. (4). The set $\Lambda = \{X_Y|Y \in \mathbf{Y}\}$ is created. Let $\Lambda_{\min} = \{X|X \text{ is a lower boundary vector in } \Lambda\}$. Then,

Lemma 2 *Λ_{min} is the set of d-MPs for d.*

2.2 Evaluation of Network Reliability

Given a demand d, the reliability denoted by ω_d is the probability at sink node that the maximal flow in the network is no less than d, i.e., $\omega_d \equiv \Pr\{X|\varpi_X \geq d\}$. To calculate ω_d, find the lower boundary vectors directly in the set $\{X|\varpi_X \geq d\}$. A lower boundary vector X is said to be a d-MP for d if and only if (i) $\varpi_X \geq d$ and (ii) $\varpi_W < d$ for any other vector W such that $W < X$, in which $W \leq X$ if and only if $w_j \leq x_j$ for each $j = 1, 2, \ldots, n$ and $W < X$ if and only if $W \leq X$ and $w_j < x_j$

for at least one j. Suppose there are totally q d-MPs for d: $X_1, X_2, \ldots, X_q$. The reliability is equal to

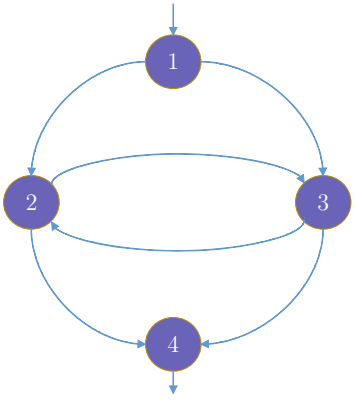
$$\omega_d = \Pr\left\{\bigcup_{k=1}^{q}\{X|X \geq X_k\}\right\}, \tag{5}$$

which can be calculated by reduced recursive inclusion-exclusion principle (RRIEP) [5].

3 Network Representation

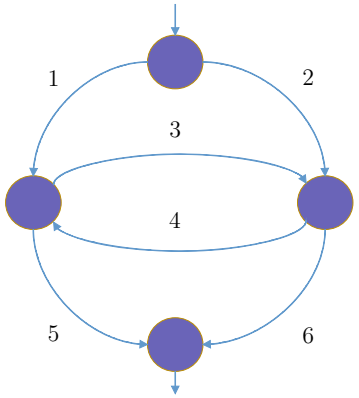
In investigating network problems, the method for expressing network structure is important. Usually, the adjacency matrix [25] (AM) is employed. An adjacency matrix is a square matrix used to represent a finite graph. The elements of the matrix indicate whether pairs of vertices (arcs) are adjacent or not in the graph. Figure 1 gives an example of node-oriented AM expression, where the entry in the matrix indicates the number of links between 'from' node and 'to' node. Such entry implicitly indicates the direction of the links between two nodes. Thus, multiple same direction links between two nodes are allowed in such AM expression. Figure 2 gives another arc-oriented AM expression. However, only 1 or 0 are allowed in this expression. Since both nodes-oriented and arc-oriented AM are equivalent in nature, we will take node-oriented AM in the following discussion.

A network is a graph except that the network has source nodes and sink nodes. Therefore, using AM to solve network problems normally faces difficulties in handling source nodes and sink nodes. Furthermore, in the view points of computer science, AM is not the best choice for program implementation. Specially, in solving network reliability, the linked path structure (LPS) [10] is now more and more



Fm\To	1	2	3	4
1	0	1	1	0
2	0	0	1	1
3	0	1	0	1
4	0	0	0	0

Fig. 1 A graph and its node-oriented adjacency matrix expression



Fm\To	1	2	3	4	5	6
1	0	0	1	0	1	0
2	0	0	0	1	0	1
3	0	0	0	1	0	1
4	0	0	1	0	1	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Fig. 2 A graph and its arc-oriented adjacency matrix expression

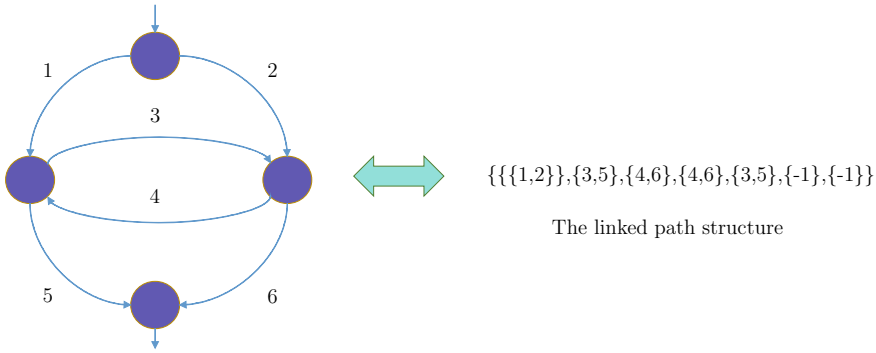


Fig. 3 A network and its linked path structure

popular in the literature. A LPS is a hybrid form of node-oriented and arc-oriented representation of a network. In LPS, entry 0 indicates the source nodes, and the negative entry values denote the sink nodes. The other entry values in LPS are vectors representing nodes with outbound arcs as their values in the vector. The index of LPS is the arc no. pointing the node (i.e. the entry value). Therefore, a LPS gives all the necessary information for networks in the applications. Figure 3 shows a LPS example. This LPS has a source node {1,2} with outbound arc 1 and 2. Arc 1 is connected with node {3,5} with outbound arc 3 and 5. Arc 2 is connected with node {4,6} with outbound arc 4 and 6. Arc 3 is connected with node {4,6} with outbound arc 4 and 6. Arc 4 is connected with node {3,5} with outbound arc 3 and 5. Arc 5 is connected with sink node $\{-1\}$. Arc 6 is connected with sink node $\{-1\}$. Thus, a tool for transforming AM to LPS is required.

An AM is a $N \times N$ square matrix V, where v_{ij} represents the number of arcs from node i to node j. We have the following lemas hold.

Lemma 3 $\exists i$, $\sum_{j=1}^{N} v_{ij} = 0$, *then, node* i *is a sink node.*

Lemma 4 $\exists j$, $\sum_{i=1}^{N} v_{ij} = 0$, *then, node* j *is a source node.*

Lemma 5 $\exists i$, j *and* $i \neq j$, $v_{ij} = v_{ji}$, *then, node* i *and* j *have undirected links.*

By above lemas, we can derive the following novel approaches for transforming AM to LPS.

3.1 The Proposed Approach

Let L be a LPS for the network. Then, we have the following properties.

Property 1 $\exists l_k \in L$, then, l_0 is the set of source nodes.

Property 2 $\exists l_k \in L$ and $k > 0$, then, l_k is a vector of arcs outbound from the node connected by arc k.

Property 3 $\exists l_k \in L$ and $k > 0$, if l_k is a vector of one negative integer, then, arc k connected with a sink node.

3.2 Algorithm of Transformation from AM to LPS

This algorithm will create two outputs: the LPS L, and a list of arc-node pairs K for verification purposes.

Algorithm 1

1. Input $N \times N$. // the AM for the underlined network.
2. $k = 1$ and $R = \phi$ // k the index for LPS and R the set of outbound arcs for the nodes.
3. For $i \in \{1, \ldots, N\}$ do // column handling.
4. For $j \in \{1, \ldots, N\}$ do // row handling.
5. if $v_{ij} - 1 < 0$ then continue,
6. else $W = W \cup \{k, \{i, j\}\}$, $R_i = R_i \cup \{k\}$, $k = k + 1$, and $v_{ij} = v_{ij} - 1$, goto step 5.

endfor
endfor
7. $k = 0$
8. For $i \in \{1, \ldots, N\}$ do // search for sink nodes.
9. if $R_i = \phi$, then, $k = k - 1$ and $R_i = \{k\}$.

endfor
10. $k = 0$

11. For $j \in \{1, \ldots, N\}$ do // search for source nodes.
12. if $\sum_{i=1}^{n} v_{ij} = 0$, then, $L_0 = L_0 \cup R_j$.
endfor
13. For $k \in \{1, \ldots, z\}$ do // search for LPS.
14. $L_k = R_{W_k(3)}$. // $W_k(3)$ means the third value in W_k.
endfor
15. Return.

The complexity analysis for this algorithm is as follows. For the AM transformation, it requires $O(N^2)$ to do step 3. Therefore, the total complexity is $O(N^2)$.

3.3 Illustrative Examples

Figure 4 gives five examples for this illustration, where Fig. 4 (a) is a 9×12 (nodes $\times$ arcs) network, (b) is a 16×20 network, (c) is a 25×36 network, (d) is a 32×48 network, and (e) is a 32×48 network. The transformed LPS are listed in Table 1.

Table 1 shows the LPS transformed correspondingly. All the cases are running on a PC with intel CPU of Core i5 6200U 2.30 GHz 2.40 GHz and 8 GB RAM.

4 Searching for MPs

Firstly, we transform the undirected edge to its directed form of arcs, and to search for all MPs in that transformed network. Then, the directed MPs are transformed back to the original form of the MPs. The following gives the details.

4.1 The General Form of a Network Transforming to Its Directed One

The difference between the general form of a network and its directed one is that there is only one way of flow for an arc in the directed form, but its general form may have both directions of flow for an edge. The following properties describe the details.

Property 4 An edge from a source node in a general flow network only has outgoing flow.

Property 5 An edge to a sink node in a general flow network only has ingoing flow.

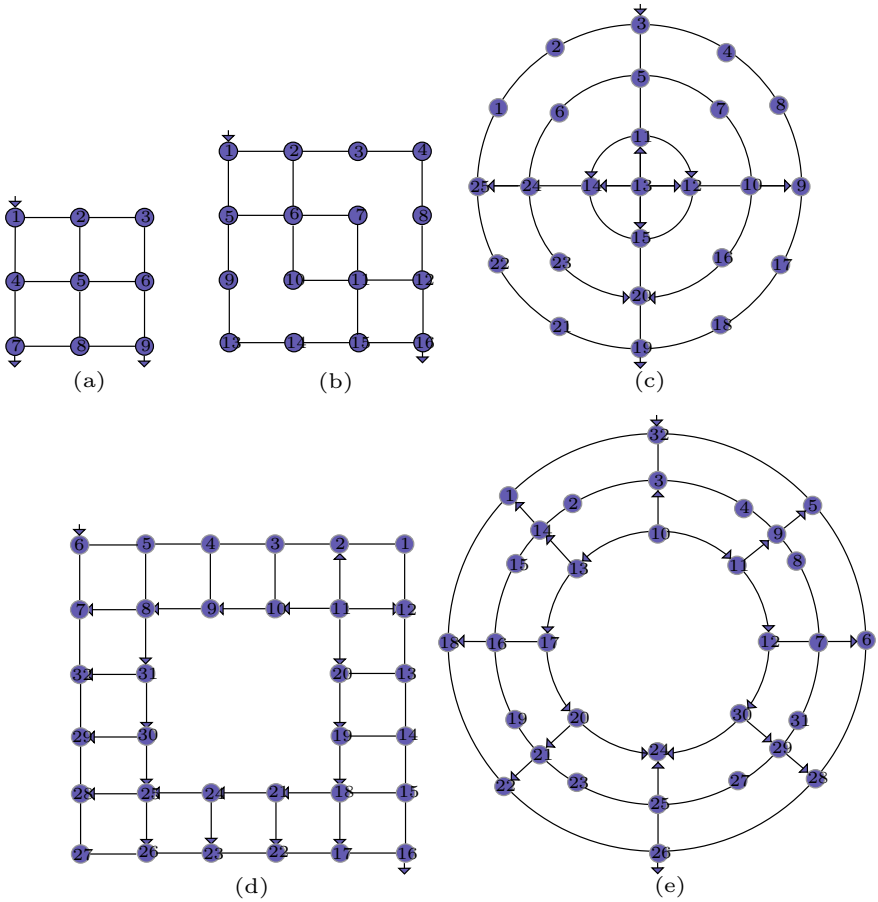


Fig. 4 Five examples for AM transformation

Property 6 A pair of reverse arcs is equivalent to a single undirected edge.

Property 7 The transformed network has an arc number greater than the total number of edges in the original network.

Thus, the edges from the source nodes or to the sink nodes have only one way of flows. The undirected edges on the other part of the network can be transformed to two reverse arcs. Figure 5 illustrates the transformation. Note that the additional reverse arcs are required for modeling the reverse flows. The backward transformation of the directed MP is to replace the added arc number in the MP back to its counterpart.

The following property shows that a cycle of length two for the undirected or hybrid edges is a parallel. Figure 6 illustrates the situation.

Table 1 The results derived by the algorithm for the five examples

Examples	LPS
(9×12)	{{(1, 2)}, (3, 4), (7, 8), (5, 6), (9, 10, 11, 12), (3, 4), (13, 14, 15), (9, 10, 11, 12), (16, 17), (3, 4), (7, 8), (13, 14, 15), (18, 19, 20), (5, 6), (9, 10, 11, 12), (−1), (7, 8), (18, 19, 20), (9, 10, 11, 12), (16, 17), (−1)}
(16×20)	{{(1, 2)}, (3, 4), (9, 10), (5, 6), (11, 12, 13, 14), (3, 4), (7, 8), (5, 6), (17, 18), (11, 12, 13, 14), (19, 20), (3, 4), (9, 10), (15, 16), (21, 22), (11, 12, 13, 14), (23, 24, 25, 26), (7, 8), (27, 28, 29), (9, 10), (30, 31), (11, 12, 13, 14), (23, 24, 25, 26), (15, 16), (21, 22), (27, 28, 29), (34, 35, 36), (17, 18), (23, 24, 25, 26), (−1), (19, 20), (32, 33), (30, 31), (34, 35, 36), (23, 24, 25, 26), (32, 33), (−1)}
(25×36)	{{(4, 5, 6), (27, 28, 29, 30)}, (3), (54, 55), (1, 2), (3), (7), (8, 9), (14, 15), (10, 11), (12, 13), (8, 9), (50, 51, 52, 53), (8, 9), (18, 19, 20, 21), (7), (16, 17), (14, 15), (38, 39), (12, 13), (16, 17), (25, 26), (36, 37), (8, 9), (25, 26), (31, 32), (18, 19, 20, 21), (33, 34, 35), (22, 23, 24), (25, 26), (31, 32), (33, 34, 35), (33, 34, 35), (50, 51, 52, 53), (25, 26), (31, 32), (42, 43), (18, 19, 20, 21), (42, 43), (16, 17), (40, 41), (38, 39), (−1), (33, 34, 35), (−1), (−1), (46, 47), (44, 45), (54, 55), (42, 43), (50, 51, 52, 53), (10, 11), (31, 32), (48, 49), (54, 55), (1, 2), (46, 47)}
(32×48)	{{(13, 14), (23, 24, 25, 26)}, (3, 4), (27, 28), (1, 2), (5, 6, 7), (3, 4), (8, 9, 10), (21, 22), (5, 6, 7), (11, 12), (19, 20), (8, 9, 10), (16, 17, 18), (11, 12), (15), (68, 69), (11, 12), (15), (66, 67), (8, 9, 10), (16, 17, 18), (5, 6, 7), (19, 20), (3, 4), (21, 22), (27, 28), (44, 45), (1, 2), (29, 30), (27, 28), (44, 45), (29, 30), (34, 35, 36), (42, 43), (31, 32, 33), (−1), (39, 40, 41), (−1), (48, 49), (34, 35, 36), (37, 38), (46, 47), (31, 32, 33), (39, 40, 41), (29, 30), (42, 43), (48, 49), (52, 53), (37, 38), (50, 51), (48, 49), (56, 57), (50, 51), (54, 55), (56, 57), (60, 61), (50, 51), (58, 59), (56, 57), (60, 61), (58, 59), (62, 63), (60, 61), (68, 69), (54, 55), (62, 63), (64, 65), (68, 69), (15), (62, 63)}
(32×48)	{{(20, 21, 22), (68, 69, 70)}, (40, 41), (4, 5), (29, 30, 31), (2, 3), (6, 7), (4, 5), (17, 18, 19), (9, 10), (8), (57, 58), (9, 10), (15, 16), (25, 26), (66, 67), (11, 12, 13, 14), (17, 18, 19), (6, 7), (8), (15, 16), (4, 5), (23, 24), (27, 28), (17, 18, 19), (25, 26), (11, 12, 13, 14), (64, 65), (29, 30, 31), (38, 39), (1), (2, 3), (32, 33), (29, 30, 31), (34, 35, 36, 37), (32, 33), (38, 39), (40, 41), (42, 43), (34, 35, 36, 37), (44, 45), (1), (49, 50), (34, 35, 36, 37), (46, 47, 48), (46, 47, 48), (−1), (42, 43), (49, 50), (51, 52), (40, 41), (−2), (46, 47, 48), (53, 54, 55, 56), (51, 52), (−1), (−2), (57, 58), (53, 54, 55, 56), (61, 62, 63), (9, 10), (−2), (57, 58), (59, 60), (66, 67), (−1), (61, 62, 63), (11, 12, 13, 14), (61, 62, 63), (1), (4, 5), (8)}

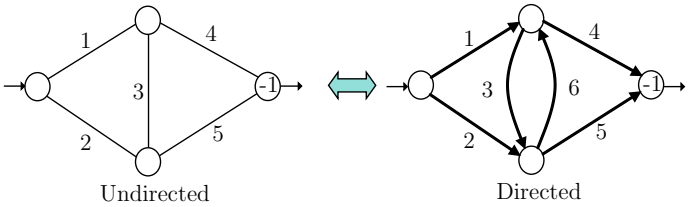
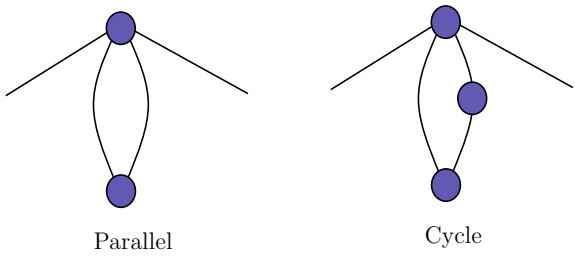


Fig. 5 The transformation between undirected and directed networks

Fig. 6 The parallel and cycle for an undirected network



Property 8 A cycle of length two for the undirected or hybrid edges is a parallel.

So, the minimal cycle length in a general flow network is 3 as we define it. This means that the minimal cycle length in its corresponding directed form should be set to no less than 3.

From the above notations, we have the following lemmas.

Lemma 6 (Cycle detection) *$\rho_j \cup v$ constructs a cycle iff $\exists v_k \in \rho_j$ such that $v_k = v$.*

From Lemma 6, we can search for all cycles in the network. Because a loop is a special case of a cycle, Lemma 6 lets such networks be applicable. Let S and T denote the source nodes and sink ones.

Lemma 7 (Minimal Path detection) *ρ is a MP iff ρ has no cycles, and the corresponding $w = s \cup v_1 \cup v_2 \cup \cdots \cup v_k \cup t$, where $v_k \in B - (S \cup T)$, $s \in S$, and $t \in T$.*

4.2 Algorithm for Searching MPs

Given the reverse arc information for all undirected links, the algorithm is given in the form in which backtracking is indicated at the line of activation.

Algorithm 2 Searching for all MP.

1. Input L. // Input the direct form of network configuration.
2. For $v \in L(0) = S$, and $v \neq \phi$, do the following steps.// Select a candidate.
3. $R = v \cup L - S$. // Construct the working list R with starting index 0.
4. Let $j = 0$, $\rho = w = \phi$. // Initialize the starting point, the current path, node path.
5. If $j < 0$, then $\rho = \cup_{2 \leq l}\{\rho(l)\}$, transform back to undirected form if necessary, output the MP, and backtrack. // Got an MP.
6. Else, begin $v = R(j)$. // Go to the node pointed by j.
7. If $w(k) = v$ is true, then backtrack.
8. Else, if $j \in v$, and $\{j\} \neq \phi$, then $\rho = \{j\} \cup \rho$, $w = \{v\} \cup w$, and call Step 5 and its following. // Recursive call.
9. Else, backtrack. // Exhaust the outgoing branches of the node.

End.
end For. // Searching finished.

4.3 Illustrative Examples

Figure 7 shows the illustrative network in the literature [31]. The results are given in Table 2. Besides, all cycles were found too. No initial network is required for the proposed approach, and the CPU time for this case was only 0.03 s.

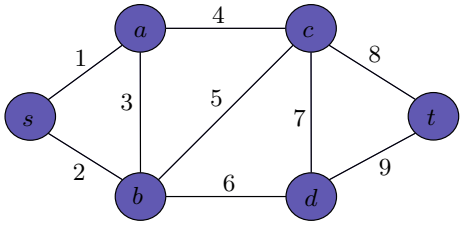


Fig. 7 The ARPA network

Table 2 The MPs searched

	The proposed approach
MPs	{2, 3, 4, 8} {2, 3, 4, 7, 9} {1, 4, 5, 6, 9} {2, 6, 7, 8} {1, 3, 6, 7, 8} {1, 3, 5, 7, 9} {1, 3, 5, 8} {1, 3, 6, 9} {1, 4, 7, 9} {1, 4, 8} {2, 5, 7, 9} {2, 5, 8} {2, 6, 9}
Cycles	{4, 5, 3} {4, 7, 6, 3} {6, 7, 5}

5 Searching for *d*-MPs

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the states of the network. The following property and lemmas hold.

Property 9 (Path existence) If a_i has a path, then, $x_i \geq 0$.

Lemma 8 (Cycle existence) *$\rho_j \cup v$ is a cycle iff $\exists v_k \in \rho_j$ with $v_k = v$.*

By Property 9 and Lemma 8, all cycles or loops in the network can be found. Yeh [29] proposed the following lemma.

Lemma 9 (*d*-MP existence) *X is a d-MP iff no cycle exists in X.*

5.1 Flow Enumerations

Based on the Chen's work [8], $\mathbf{Y}$ can be fast enumerated by optimal re-arrangement of the formulas in Ψ_Y and Eq. (3). Next, we build X_Y by $Y \in \mathbf{Y}$ with Eq. (4). Then, $\Lambda = \{X_Y | Y \in \mathbf{Y}\}$ is created for all X_Y. Finally, $\Lambda_{min} = \{X | X$ is a d-MP in $\Lambda\}$ is created. Let Ξ_{Ψ_Y} be the optimal fast enumeration form (OFE-form) of formulas in Ψ_Y [8]. The fast enumeration of flows are created from these formulas.

5.2 Algorithm for Searching d-MPs

Let $p_j = \min\{x_i | a_i \in \rho_j\}$ and L be the linked path structure for the network [10]. Based on the above subsection discussion, The following algorithm is proposed.

Algorithm 3 Searching for all d-MPs for the network.

1. Enumerate $Y = (y_1, y_2, \ldots, y_z)$ under OFE-form.
   a. **fast enumerate** $0 \leq y_j \leq p_j$ **under** Ξ_{Ψ_Y} **do**
   b. **if** y_j satisfies $\sum_{j=1}^{z} y_j = d$,
      then $\mathbf{Y} = \mathbf{Y} \cup \{Y\}$.
      end enumerate.
2. Generate Λ_{min} via $\mathbf{Y}$.
   a. **for** Y in $\mathbf{Y}$ **do**
   b. $x_i = \sum_{j=1}^{z} \{y_j | a_i \in \rho_j\}$ for $i = 1, 2, \ldots, N$.
   c. **for** $1 \leq i \leq N$ **do**
   d. **if** $x_i \geq 0$, **then** $I = I \cup \{i\}$. //Non-zero columns.
      end for.
   e. Let $\rho = \phi$.

f. **for** $i \in I$ and $i > 0$ **do**
g. **if** not($L_i \cap I \neq \phi$ or $L_i \in \rho$), **then** recursive call Step f by replacing I with L_i and $\rho = \rho \cup L_i$.
h. **else** goto Step a.
end for.
i. $\Lambda_{min} = \Lambda_{min} \cup \{X\}$.
endfor. // Searching finished.
j. **Return** Λ_{min}.

Step 1 requires $O(\prod_{k=1}^{u} r_k)$ to generate all feasible X, since OFE-form is by re-arranging the original Ψ_X into u groups of alternative orders, and r_k is the total number of enumerations in the k^{th} group [11]. Step 2 performs d-MPs searching. It's computation complexity is $O(|B| \binom{z+d}{d-1})$.

5.3 Illustrative Examples

Figure 8 gives several complicated examples for exploration. Figure 8 shows (a) 25×32 (Vertices $\times$ Arcs) network, (b) 25×36 network, (c) 32×48 network, and (d) 32×48 network. Table 3 lists the results derived by the proposed algorithm. There are 26, 28, 28 and 26 6-MPs for Fig. 8a–d respectively, and 49,907, 34,784, 344,624 and 76,713 6-MPs derived from them, respectively. The total CPU seconds for them are 40, 388.002, 45, 900.912, 70, 988.449 and 144, 412.762 s, respectively. The results denotes the efficiency for applying the proposed algorithm.

6 Calculating the Union Probability for d-MPs

The network reliability is also the union probability for d-MPs. The most popular method to calculate the probability of union events is the inclusion-exclusion principle (IEP), which originates from the idea of Abraham de Moivre (1718) [22]. For example, Fig. 9 illustrates the principle of IEP for three events, and Eq. (6) explains how to calculate the probability of $\{A \cup B \cup C\}$.

$$\begin{aligned} \Pr\{A \cup B \cup C\} = & \Pr\{A\} + \Pr\{B\} + \Pr\{C\} \\ & - \Pr\{A \cap B\} - \Pr\{B \cap C\} \\ & - \Pr\{A \cap C\} + \Pr\{A \cap B \cap C\} \end{aligned} \tag{6}$$

According to IEP, it equals to the sum of the probability of all individual events, minus the probability of the intersection of every compound events, and plus the probability of the intersection of all events. In other words, IEP is equivalent to do the power set expansion. So, it is a general principle, and can be applied to any kind of events. However, the power set expansion makes the computation complexity

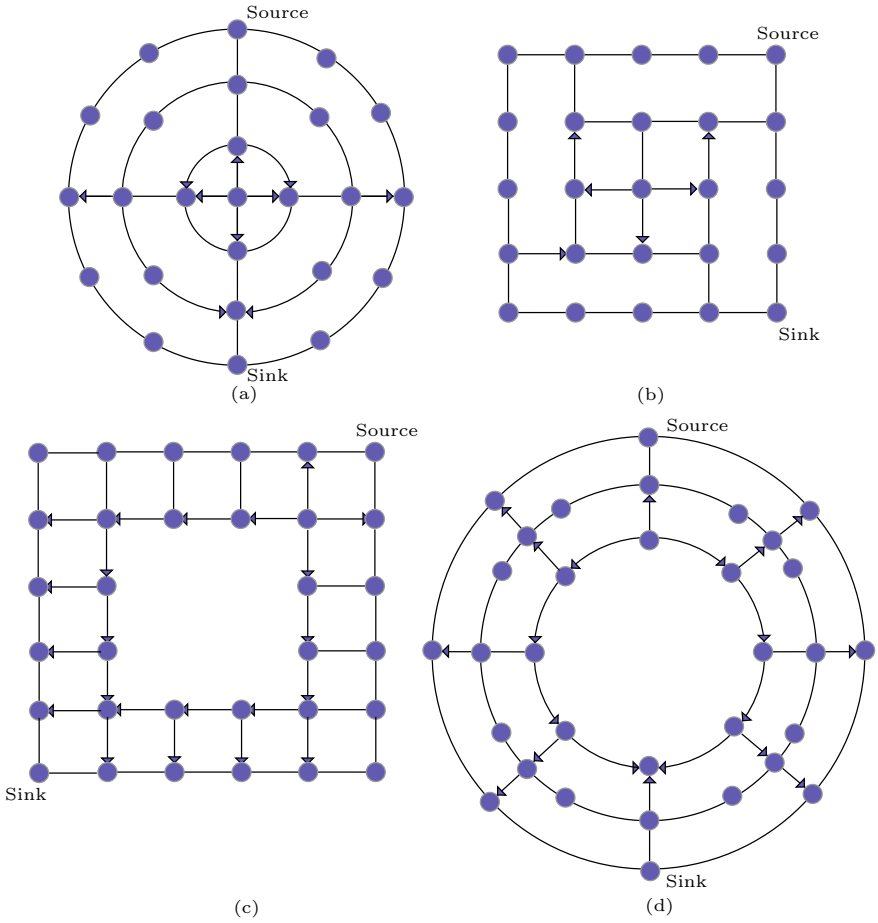
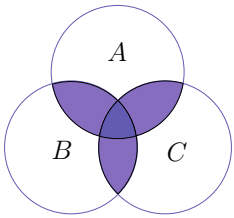


Fig. 8 Several complicated examples

Table 3 The CPU seconds for the complicated examples

(Vertices × Arcs)	# of MPs	# of 6-MPs	CPU (s)
(25 × 32)	26	49,907	40,388.002
(25 × 36)	28	34,784	45,900.912
(32 × 48)	28	344,624	70,988.449
(32 × 48)	26	76,713	144,412.762

Fig. 9 An union of 3 events



of IEP equals exactly to $O(2^Q)$ no matter what the events are. A much efficient method namely Recursive Inclusion-Exclusion Principle (RIEP) was constructed by rearranging Eq. (6) to its recursive form as Eq. (7). The computation complexity is also $O(2^Q)$ in the worse cases, but it usually has 10 times efficiency than IEP in the normal cases.

$$\begin{aligned}\Pr\{A \cup B \cup C\} &= \Pr\{A\} \\ &\quad + \Pr\{B\} - \Pr\{A \cap B\} \\ &\quad + \Pr\{C\} - \Pr\{B \cap C\} \\ &\quad - \Pr\{A \cap C\} + \Pr\{A \cap B \cap C\} \\ &= \Pr\{A\} \\ &\quad + \Pr\{B\} - \Pr\{A \cap B\} \\ &\quad + \Pr\{C\} - \Pr\{B \cap C \cup A \cap C\}\end{aligned} \tag{7}$$

This chapter presents an innovated reduction method for the recursive inclusion-exclusion principle (RRIEP) to calculate the probability of union events such that the calculation terms can be mostly eliminated, especially in the cases of events with monotonic property like network reliability applications. The network reliability is the probability of a live connection between source nodes and sink nodes. Such reliability can be calculated by means of minimal paths (MPs) [10, 12, 24] or minimal cuts (MCs) [1, 16, 30]. When the network is live, there are several minimum path vectors with respect to system state d, called d-MP, can be found. Then, the network reliability is the union probability of all these d-MPs.

6.1 Inclusion-Exclusion Principle

Let $\{A_1, A_2, \ldots, A_n\}$ be n events. Then, the probability of union events in terms of IEP is as follows.

$$\Pr\{\bigcup_{i=1}^{n} A_i\} = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=k}} \Pr\{\bigcap_{i \in I} A_i\} \right). \tag{8}$$

6.2 Recursive Inclusion-Exclusion Principle

Figure 10 presents the concept of recursive inclusion-exclusion principle. Let $\pi(X_1, X_2, \ldots, X_n) \equiv \Pr\{X_1 \cup X_2 \cup \cdots \cup X_n\}$ be the function of probability of union events. According to the equivalent rearrangement, we have the following definition of recursive inclusion-exclusion principle.

$$\begin{aligned} \pi(A_1) &= \Pr\{A_1\}, \quad \text{for } n = 1, \\ \pi(A_1, A_2, \ldots, A_n) &= \sum_{i=1}^{n} (\Pr\{A_i\} \\ -\pi(A_1 \cap A_i, A_2 \cap A_i, \ldots, A_{i-1} \cap A_i)), & \\ &\text{for } n > 1, \end{aligned} \tag{9}$$

where $\pi(\bullet)$ is a recursive function.

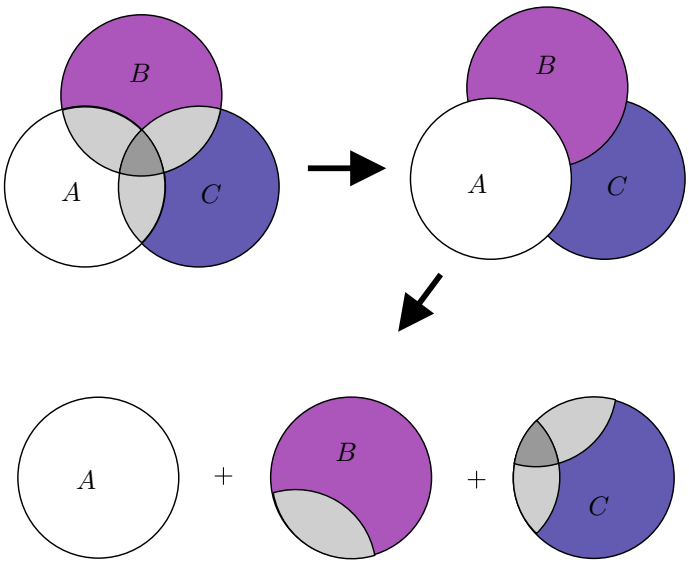


Fig. 10 An illustration of rearrangement

6.3 Reduced Recursive Inclusion-Exclusion Principle

One term is defined first.

Definition 1 (*Reduction*) $\{X_1, X_2, \ldots, X_m\}$ is a reduction of $\{Y_1, Y_2, \ldots, Y_n\}$, if for any $m, n \in \mathbb{N}$ and $m < n$ such that $\{X_1 \cup X_2 \cup \ldots \cup X_m\} = \{Y_1 \cup Y_2 \cup \ldots \cup Y_n\}$.

We have the following lemmas.

Lemma 10 *$\{X_1, X_2, ..., X_m\}$ is a reduction of $\{Y_1, Y_2, ..., Y_n\}$, then $\pi(X_1, X_2, \ldots, X_m) = \pi(Y_1, Y_2, \ldots, Y_n)$.*

Lemma 11 *Along with Lemma 10, the computation complexity of $\pi(X_1, X_2, \ldots, X_m)$ is less than that of $\pi(Y_1, Y_2, \ldots, Y_n)$.*

Let $\{B_1, B_2, \ldots, B_{m_i}\}$ be a reduction of $\{A_1 \cap A_i, A_2 \cap A_i, \ldots, A_{i-1} \cap A_i\}$. From Lemma 10, Eq. (9) can be rewritten as:

$$\begin{aligned} \pi(A_1) &= \Pr\{A_1\}, \quad \text{for } n = 1, \\ \pi(A_1, A_2, \ldots, A_n) &= \sum_{i=1}^{N} (\Pr\{A_i\} \\ -\pi(A_1 \cap A_i, A_2 \cap A_i, \ldots, A_{i-1} \cap A_i)), & \qquad (10) \\ &= \sum_{i=1}^{n} \Pr\{A_i\} \\ -\pi(B_1, B_2, \ldots, B_{m_i})), &\quad \text{for } n > 1, \end{aligned}$$

By Lemma 11, RRIEP is much efficient than RIEP.

In network reliability applications, it is easier to implement the reduction of union events. At first, we have the following property.

Property 10 (*Monotonicity*) Let $\bar{Q} = (q_1, q_2, \ldots, q_h)$ and $\bar{W} = (w_1, w_2, \ldots, w_h)$ be two capacity vectors in a network. Then, $\bar{Q}$ and $\bar{W}$ satisfy the following Monotonicity properties:

1. $\bar{Q} \leq \bar{W}$ iff $q_j \leq w_j$ for each $j = 1, 2, \ldots, h$.
2. $\bar{Q} < \bar{W}$ iff $\bar{Q} \leq \bar{W}$ and $q_j < w_j$ for at least one j.

6.4 Network Reliability Calculation

Given a demand d, the network reliability denoted by ω_d is the probability that the maximal flow is no less than d, i.e., $\omega_d = \Pr\{X | \varpi_X \geq d\}$. To calculate ω_d, it is advantageously to find the minimal vector (namely, d-MP) in the set $\{X | \varpi_X \geq d\}$.

Suppose there are totally h d-MPs for d: $X_1, X_2, \ldots, X_h$, the network reliability is equal to

$$\omega_d = \Pr\left\{\bigcup_{k=1}^{h}\{X|X \geq X_k\}\right\}, \tag{11}$$

Let $\bar{Q} \oplus \bar{W} \equiv \{g_i|g_i = \max(q_i, w_i), q_i \in \bar{Q}, w_i \in \bar{W}, \forall i\}$. The following RRIEP algorithm can be used to calculate ω_d.

Algorithm 4 RRIEP for network reliability calculation.

1. $dMP = \{X_1, X_2,\ldots,X_h\}$.
2. $\omega = 0$.
3. If $h = 0$, then return ω,
4. For $1 \leq i \leq |dMP|$ do
5. $\omega_2 = Pr\{X \geq X_i\}$.
6. For $1 \leq j < i$ do
7. $X_{j,i} = X_j \oplus X_i$.
   End for.
8. Let $E = \{X_{j,i}|1 \leq j < i\}$ and $W_l, W_k \in E$.
9. For $k \notin J$ and $1 \leq k \leq ||E||$ do // Reduction process.
10. For $l \notin J$ and $k < l \leq ||E||$ do
11. If $W_l \leq W_k$, then $J = J \cup \{k\}$ and go to Step 9.
    Else, if $W_l > W_k$, then $J = J \cup \{l\}$.
    End for.
12. $dMP^* = dMP^* \cup \{W_k\}$.
    End for.
13. Recursive call Step 1 with the new dMP^* and get the value ω_3.
14. $\omega = \omega + \omega_2 - \omega_3$.
    End for.
15. return ω.

The reduction implementation is started at Step 6.4, which greatly improves the efficiency of RRIEP algorithm. Since RRIEP can reduce the computation efforts in most of the time, the computation complexity in the worst cases still need $O(2^Q)$ in which no reduction can be performed at all.

6.5 *Illustrative Examples*

We randomly generate d-MP vectors of length 20, and compare IEP, RIEP, and RRIEP with efficiency. Table 4 gives the results. All benchmarks were implemented with PROLOG on a personal computer with the CPU of Intel Core i7-2600 CPU@ 3.40 GHz, 3.40 GHz and with 16 GB RAM.

Table 4 The comparison of CPU time (s) for benchmarks

# of d-MPs	IEP	RIEP	RRIEP
15	1.765	0.390	0.0003
18	19.891	1.891	0.0005
20	75.953	2.953	0.016
25	9256.750	9.359	0.016
30	NA*	43.343	0.015
35	NA	106.141	0.031
40	NA	416.579	0.063
45	NA	1112.687	0.078
50	NA	2144.000	0.093
55	NA	5924.610	0.110
59	NA	8362.469	0.109

*Not stopped over 1 day

7 Conclusion

This chapter presents the primitive steps in computation of network reliability, and some of the popular examples of network reliability. Although we present the most efficient method for all stages in the literature, they are subjected to change due to new method emerging. This chapter also give several numerical examples for exploration of their calculations, which help one to follow the algorithms to go through these examples.

Acknowledgements This work was supported in part by the National Science Council, Taiwan, Republic of China, under Grant No. MOST 107-2221-E-236-004-MY3.

References

1. Abel U, Bicker R (1982) Determination of all minimal cut-sets between a vertex pair in an undirected graph. IEEE Trans Reliab 31:167 – 171
2. Aggarwal KK, Chopra YC, Bajwa JS (1982) Capacity consideration in reliability analysis of communication systems. IEEE Trans Reliab 31:177–80
3. Aggarwal KK, Gupta JS, Misra KB (1975) A simple method for reliability evaluation of a communication system. IEEE Trans Commun 23:563–565
4. Aven T (1985) Reliability evaluation of multistate systems with multistate components. IEEE Trans Reliab 34:473–479
5. Chen SG (2014) Reduced recursive inclusion-exclusion principle for the probability of union events. In: 2014 IEEE international conference on industrial engineering and engineering management. Selangor, Malaysia, pp 1–3
6. Chen SG (2014) Reduced recursive sum of disjoint product in network reliability. In: 2014 the 20th ISSAT international conference on reliability and quality in design. Seattle, Washington, U.S.A., pp 170–173
7. Chen SG (2015) Search for all MCs with backtracking. Int J Reliab Qual Perform 6(2):101–106

8. Chen SG (2016) Optimal re-arrangement in fast enumeration for integer programming problems. In: 2016 IEEE international conference on industrial engineering and engineering management, pp 1255–1258
9. Chen SG (2019) A way from adjacency matrix to linked path structure. In: Proceedings of the 25th ISSAT international conference on reliability and quality in design, pp 20–23
10. Chen SG, Lin YK (2012) Search for all minimal paths in a general large flow network. IEEE Trans Reliab 61(4):949–956
11. Chen SG, Lin YK (2016) Searching for d-MPs with fast enumeration. J Comput Sci 17:139–147
12. Colbourn CJ (1987) The combinatorics of network reliability. Oxford University Press, UK
13. Doulliez P, Jamoulle J (1972) Transportation networks with random arc capacities. Recherche Operationnelle 3:45–60
14. Ford LR, Fulkerson DR (1962) Flows in networks. Princeton University Press, NJ
15. Jane CC, Lin JS, Yuan J (1993) On reliability evaluation of a limited-flow network in terms of minimal cutsets. IEEE Trans Reliab 42:354–361, 368
16. Jasmon GB, Foong KW (1987) A method for evaluating all the minimal cuts of a graph. IEEE Trans Reliab 36:538 – 545
17. Lee SH (1980) Reliability evaluation of a flow network. IEEE Trans Reliab 29:24–26
18. Lin JS, Jane CC, Yuan J (1995) On reliability evaluation of a capacitated-flow network in terms of minimal pathsets. Networks 25:131–138
19. Lin YK (2001) On reliability evaluation of a stochastic-flow network in terms of minimal cuts. J Chin Inst Ind Eng 18:49–54
20. Lin YK (2001) A simple algorithm for reliability evaluation of a stochastic-flow network with node failure. Comput Oper Res 28(13):1277–1285
21. Lin YK, Chen SG (2019) An efficient searching method for minimal path vectors in multi-state networks. Ann Oper Res. 1–12. https://doi.org/10.1007/s10479-019-03158-6
22. Roberts FS, Tesman B (2009) Applied Combinatorics, 2nd edn. CRC Press
23. Schrijver A (2002) On the history of the transportation and maximum flow problems. Math Program 91(3):437–445
24. Shen Y (1995) A new simple algorithm for enumerating all minimal paths and cuts of a graph. Microelectron Reliab 35:973–976
25. Wilson RJ (1972) Introduction to graph theory. Prentice-Hall
26. Xue J (1985) On multistate system analysis. IEEE Trans Reliab 34(4):329–337
27. Yeh WC (2001) A simple algorithm to search for all d-MPs with unreliable nodes. Reliab Eng Syst Saf 73:49–54
28. Yeh WC (2001) A simple approach to search for all d-MCs of a limited-flow network. Reliab Eng Syst Saf 71:15–19
29. Yeh WC (2002) A simple method to verify all d-minimal path candidates of a limited-flow network and its reliability. Int J Adv Manuf Technol 20(1):77–81
30. Yeh WC (2006) A simple algorithm to search for all MCs in networks. Eur J Oper Res 174:1694–1705
31. Yeh WC (2007) A simple heuristic algorithm for generating all minimal paths. IEEE Trans Reliab 56(3):488–494
32. Zuo MJ, Tian Z, Huang HZ (2007) An efficient method for reliability evaluation of multistate networks given all minimal path vectors. IIE Trans 39:811–817

Shin-Guang Chen received his B.Sc. in Computer Engineering, NCTU, Taiwan and his Ph.D. in IE, NCTU, Taiwan. He has been with the Institute of Industrial Management at Tungnan University, Taiwan, where he holds the rank of Distinguished Professor. His research interests are in management, reliability, ERPS and sustainable energy technologies. His recent work involves problems related to the analysis of flow networks. He has published papers in IJPR, IJIE, ESWA, JORSJ, CSTM, IJOR, JCIIE, IJRQP, APEN, IJPE, Omega, IEEE T Reliab, APM, JIMO, APJOR, AOR, etc. He is a member of IEEE, IEEE Reliab, POMS, CIIE, CERPS and CSQ.

Integrating Sentiment Analysis in Recommender Systems

Bui Thanh Hung

Abstract Customer product reviews play an important role in the customer's decision to purchase a product or use a service. Providing a useful suggestion of products to online users to increase their consumption on websites is the goal of many companies nowadays. In this paper, we propose a recommender system based on sentiment analysis. The system is built by integrating sentiment analysis to recommender system in order to generate the most accurate. We use hybrid deep learning method CNN-LSTM for sentiment analysis based on vector of words in the customer product reviews. The result in the sentiment analysis is used to combine the neighbor's item ratings to produce a prediction value for the target user. This helps the recommender system to generate efficient recommendations for that user. We do experiment in Amazon food review dataset. The proposed model shows interesting results on the impact of integrating sentiment analysis in the recommender systems.

Keywords Sentiment analysis · Word embeddings · CNN-LSTM · Matrix factorization · Recommender system

1 Introduction

With the explosive growth of the Internet, these days consumers could easily share their feedback as well as experiences with brands from different companies. This means that by collecting and analyzing the above customer feedback and mining the users' opinions, these companies can discover consumer attitudes towards them, and thus make necessary and prompt adjustments. Understanding the meaning of the reviews and correctly classify its helpfulness. The results of the classification can be used for further application such as recommender system to provide a useful suggestion of products to online users. For the sentiment analysis, there are a variety of approaches to address this, such as machine learning approach or word-based

B. T. Hung (✉)
Data Analytics & Artificial Intelligence Laboratory, Engineering - Technology Faculty, Thu Dau Mot University, 6 Tran van on Street, Phu Hoa District, Thu Dau Mot City, Binh Duong Province, Vietnam
e-mail: hungbt.cntt@tdmu.edu.vn

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_8

approaches. There are also several levels in opinion mining, such as document-level, sentence-level and aspect-based [1–3]. In this paper, we focus on the document-level of opinion mining, which means to determine whether an opinion expressed is positive, negative or neutral. This could be done through training hybrid deep learning models. In this research, at the first step, we classify the usefulness of each review using hybrid deep learning models CNN-LSTM [4]. We generate a vector representation for each customer product review, use those vectors to train our model, and evaluate the performance of our classification model. After getting the result in the first step, we integrate this in the recommender system based on user-item-rating data. We do experiment on Amazon food reviews and evaluate the performance of our model based on their RMSE score. The remainder of this paper is organized as follows: Sect. 2 introduces related work. Section 3 describes in details the proposed model. Section 4 shows the experiments, as well as discussion related to the results. Finally, Sect. 5 summarizes our work, as well as future directions.

2 Related Works

A recommender system provides suggestions to users, in multiple contexts. For example, when choosing between multiple items or providing the customer with suggested products. Most of e-commerce websites use recommender systems, where the system displays a list of recommended items to the end user. Then, identifying potentially useful items for users is the core function of a recommender system. In order to predict these, a recommender system has to be able to predict the utility of these items. Then, based on the results, the system decides which items to recommend [5, 6].

Recommendation models are mainly categorized into collaborative filtering, content-based recommender system and hybrid recommender system based on the types of input data. A collaborative filtering system suggests user items similar to those he preferred or liked in the past. A content-based recommender system suggests user items that people with similar preferences liked in the past [6, 7].

There are many researches focus on this topic. In content-based approach, almost researchers used Vector Space Model such as Term Frequency Inverse Document Frequency (TF/IDF) or Probabilistic models such as Naıve Bayes Classifier, Decision Trees or Neural Networks to model the relationship between different documents within a corpus. In collaborative filtering approach, the technique can be divided into two categories: memory-based and model-based [5–10]. Figure 1 shows recommendation techniques.

Our research is different with the others by integrating sentiment analysis in recommender systems. The hybrid deep learning approach CNN-LSTM for sentiment analysis combines of the superior features of CNN and LSTM models. And integrating sentiment analysis helps to improve quality of recommender systems.

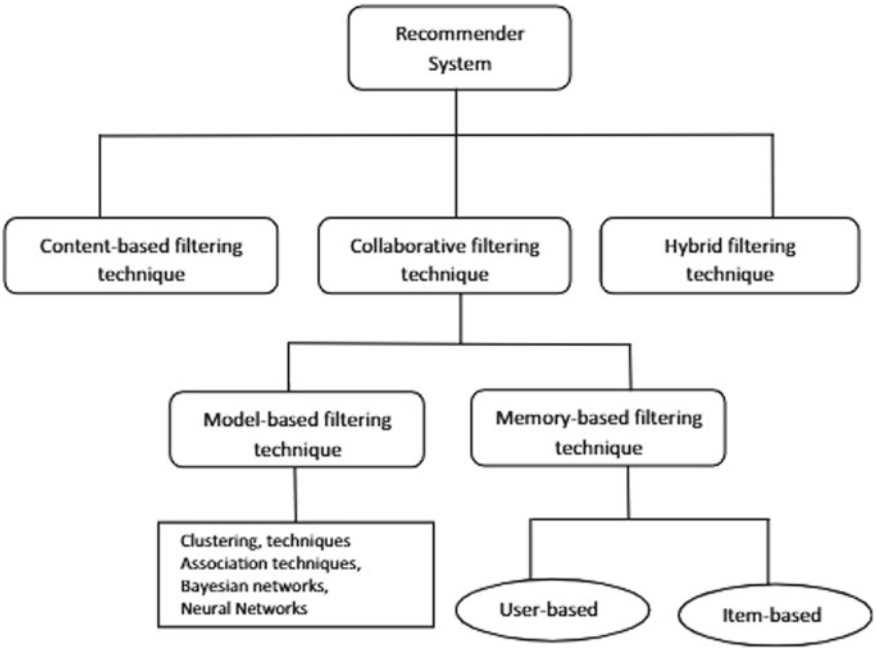
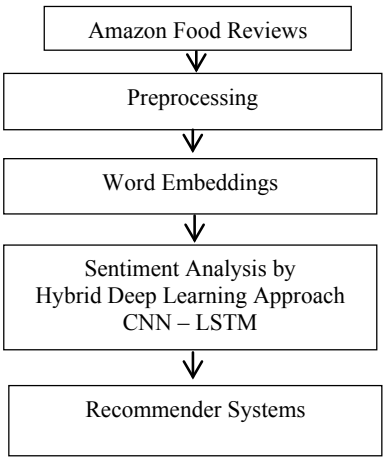


Fig. 1 Recommendation techniques

3 The Proposed Model

An overview of our architecture is shown in Fig. 2. We will describe more detail about functionality and responsibility of each layer in our model.

Fig. 2 The proposed model



3.1 Word Embeddings

Every word has reflected the structure of the word regarding the semantical/morphological/context/hierarchical/etc. information. The idea of Word embedding is to capture with them as much as possible and convert it to vectors.

Although it has been discussed long time ago the idea of using vector representation for words, it has just become popular recently the interest in word embedding, techniques that map words to vectors. Tomas Mikolov's Word2vec algorithm [11, 12] is one driver for this which uses a large amount of text to create high-dimensional (50–300 dimensional) representations of words; this captures relationships between words unaided by external annotations. It seems to capture many linguistic regularities with such representation.

In this research, we use word embedding—Continuous Bag of Words (CBOW) [11, 12]. In the continuous bag of words model, context is represented by multiple words for a given target words. For example, we could use cat and on as context words for climbed as the target word. This calls for a modification to the neural network architecture. The modification, shown below, consists of replicating the input to hidden layer connections C times, the number of context words, and adding a divide by C operation in the hidden layer neurons (Fig. 3).

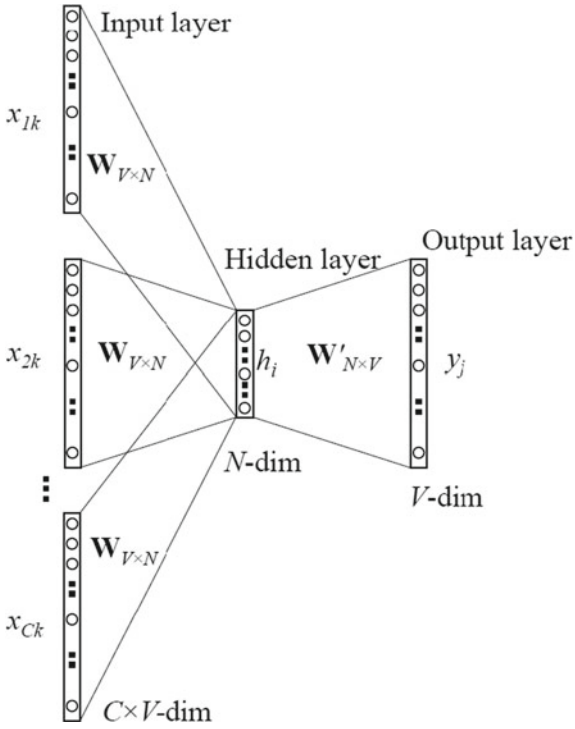


Fig. 3 Continuous Bag of Words (CBOW)

3.2 Hybrid Deep Learning CNN-LSTM

Unlike task-specific algorithms, deep learning is a member of wider family of machine learning methods which is based on learning data representations. Learning could be classified into three categories including supervised, semi-supervised or unsupervised. There are two most well-known types of deep neural networks which are Recurrent neural networks (RNNs) and Convolutional neural networks (CNNs) [13, 14]. In this section, we will describe our hybrid deep learning method by combining Convolutional neural networks with Long short term memory (CNN-LSTM) [4].

LSTM

Generally, a recurrent neural network is a type of advanced artificial neural network. RNNs can use their internal state (memory) to process sequences of inputs. RNNs have shown great successes in many natural language processing tasks. They connect previous information to present task, such as predict the next word in the sentence "clouds on the sky". Since the gap between the relevant information and the place that it needs is small, RNNs can learn to use the past information. However, in terms of Long-Term dependencies, RNNs has not yet worked well. Long short term memory [14] is a modification of the Recurrent neural networks (RNN). The main highlight of the LSTM when compared with the regular feed forward neural network is that they are able to retain the knowledge about previous outputs. The retention part is due to the feedback loop present in their architecture. Figure 4 shows a diagram of a simple LSTM cell. Individual cells are combined together to form a large network thereby befitting the term deep neural networks. The cell unit represents the memory. This cell is composed of five main elements: an input gate i, a forget gate f, an output gate o, a recurring cell state c and hidden state output h.

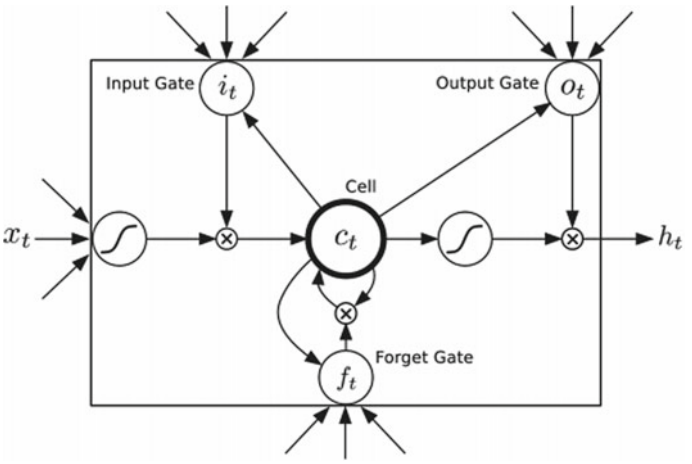


Fig. 4 The long short term memory cell

CNN

CNN includes many major architectural components which are paired with each other in multi-story structure that is: Convolution, Pooling, Activation function (ReLu, Sigmoid, Tanh), and Fully connected [13]. This model is shown in Fig. 5.

CNN-LSTM

The CNN model focused on extracting spatial information as features, whereas the LSTM model focused on extracting temporal information (from the past to the future). CNN-LSTM model, therefore, can combine and use both types of information together. The basic idea is to use the CNN model first to extract the spatial information. Then, instead of flatten them, we feed them directly to the LSTM model. Figure 6 shows hybrid CNN-LSTM model [4].

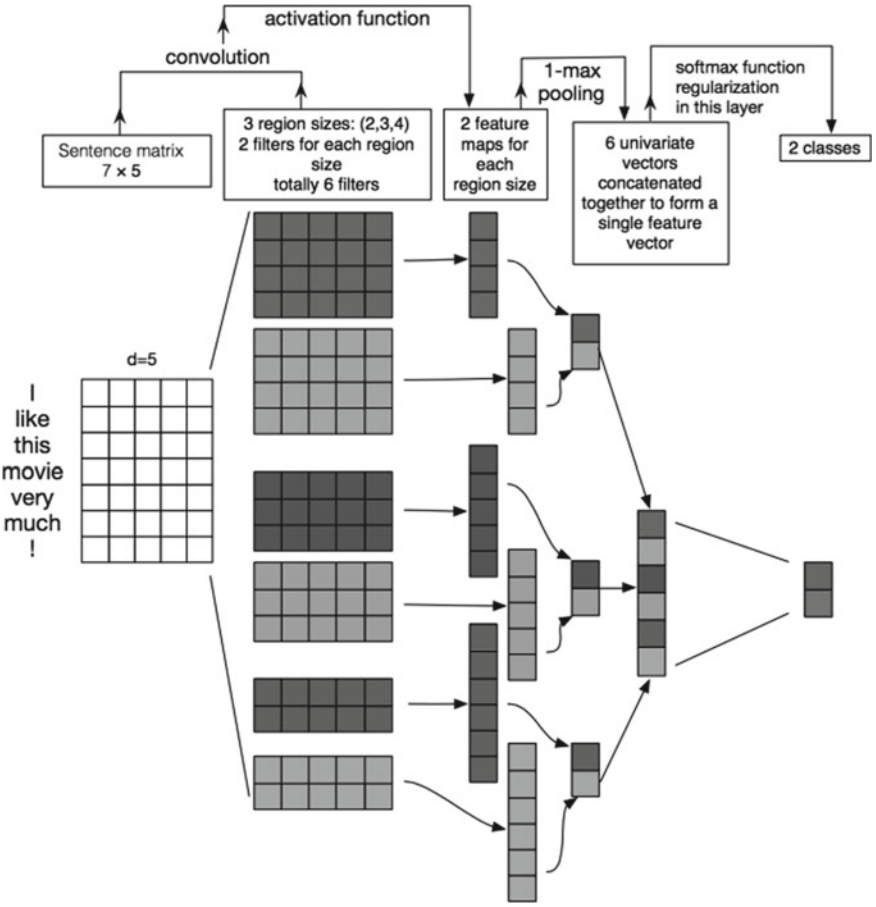


Fig. 5 CNN model

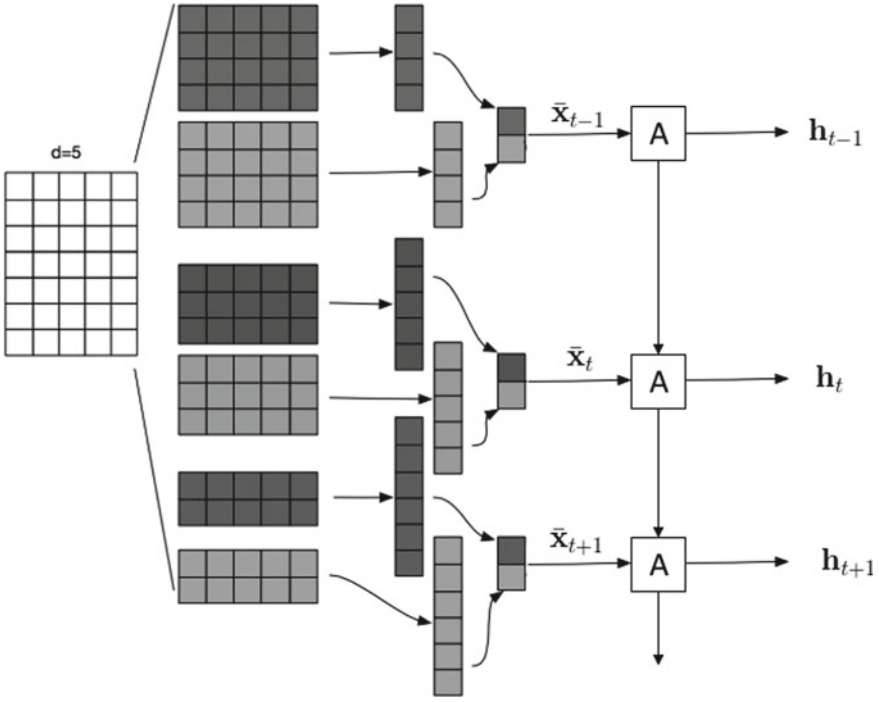


Fig. 6 CNN-LSTM

3.3 Sentiment Analysis by CNN-LSTM

The proposed model consists of five primary components: word embedding features, Convolution layer, Max-pooling Layer, Sequential Layer and Decoder. For each word embedding, CNN model extract useful affective features through a convolutional layer and max pooling layer. Such local features are then sequentially integrated using LSTM to build a text vector for sentiment analysis prediction. With the text vector learned from the CNN model, the output layer defined as:

$$\mathrm{y} = (W_d x_t + b_d) \tag{1}$$

where,

x_t	the text vector learned from the CNN models,
y	the degree of keyword of the target text,
W_d, b_d	the weight and bias associated with the CNN model.

By minimizing the mean squared error between the predicted sentiment class and actual sentiment class, the LSTM model is trained. The loss function is defined as follows:

$$L(X, y) = \frac{1}{2n} \sum_{k=1}^{n} \binom{n}{k} \left\| h(x^i) - y^i \right\|^2 \tag{2}$$

$X = \{\boldsymbol{x}^1, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{\boldsymbol{m}}\}$: a training set of text matrix$y = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^{\boldsymbol{m}}\}$: a sentiment ratings set of training sets.

3.4 Recommender System

Once the neighbors are obtained and their associated reviews are classified, a weighted average is used to combine the neighbor's item ratings to produce a prediction value for the target user. We use matrix factorization model, one of the most successful and widely used in recommender system to recommend items to user, it characterizes both items and users by vectors of factors inferred from item rating patterns.

Denote the vector for user i as q_i, and vector for product j as v_j, our prediction of the rating of user i on product j can be written as:

$$\widetilde{r}_{ij} = \dot{v}_j^T q_i \tag{3}$$

Therefore, to learn the optimal user and product vectors, our objective function is:

$$\min_{v^*, q^*} \sum_{(i,j)} \left(r_{i,j} - \dot{v}_j^T q_i \right)^2 + \left(\lambda \|v\|^2 - \|q\|^2 \right) \tag{4}$$

4 Experiments

We used Amazon food reviews as dataset [15]. All data in one database with 568,454 food reviews, this data is reviewed by customers from October 1999 to October 2012. More detail about this dataset is described in Table 1.

To evaluate the first step—sentiment analysis, we used accuracy score defined as,

Table 1 Dataset

Time	Oct 1999–Oct 2012
Reviews	568,454
Users	256,059
Products	74,258
Users with > 50 reviews	260

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where,

TP is the true positive,
TN is the true negative,
FP is the false positive,
and *FN* is the false negative.

We divide the data into training and testing dataset with ratio 7:3. We labeled for each review sentence by 0 with the rating score between 0 to 3, and 1 with the others. We used NLTK and Beautiful Soup for preprocessing. The pre-trained fast Text (Facebook AI research) is used as word embeddings in our model. For our training models, we use Keras deep learning tools which provide many useful layer and parameter. The model of the Keras suite has been trained with the following values of its parameters:

CNN: number of convolution layer nodes: 128, Kernel size: 5, Max pool: 4D
LSTM: Number of hidden nodes: 128.

We used Drop out: 0.4, Activate function in the output layer: Sigmoid, Training iteration (epochs): 1000, Batch size: 500, Optimization function: Adam, Loss function: Binary cross entropy.

We compared our result in the first step with CNN, LSTM model separately as shown in Table 2. This result shows that our proposed model—hybrid deep learning approach CNN-LSTM gets the best results.

When we got the results from sentiment analysis task, we integrated this result in the recommender systems. We evaluated this task by Root Mean Square Error (RMSE) is given by Cotter et al. as,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{u,i}(p_{u,i} - r_{u,i})^2} \tag{6}$$

where,
$p_{u,i}$ is the predicted rating for user u on item i, $r_{u,i}$ is the actual rating, n is the total number of ratings on the item set.

Root Mean Square Error (RMSE) puts more emphasis on larger absolute error and the lower the RMSE is, the better there commendation accuracy.

Table 2 The results in sentiment analysis

Model	Accuracy
CNN	0.8276
LSTM	0.8254
CNN-LSTM	0.8345

Table 3 Performance of different recommender systems

Model	RMSE
Popular(baseline)	1.7364
Matrix Factorization integrated Sentiment Analysis	1.1536

Our baseline model for recommender system is recommending popular product overall. The result is shown in Table 3. As we can see from above, matrix factorization integrated Sentiment Analysis beats the baseline model. Matrix factorization is the best-performing score comparing with the baseline.

5 Conclusion

In this paper, we experimented our model by integrating sentiment analysis in the recommender systems. We used hybrid deep learning approach CNN-LSTM for sentiment analysis task and matrix factorization for recommender systems. The result shows that our proposed model yields the highest result in both sentiment analysis and recommender systems. In future work, we would like to apply deep network to learn a function that maps item content features to corresponding matrix factorization latent factors.

References

1. Hung BT (2018) Domain-specific vs. general-purpose word representations in sentiment analysis for deep learning models. In: The 7th international conference on frontiers of intelligent computing: theory and applications, FICTA
2. Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E (2017) Learning word representations for sentiment analysis. Cogn Comput 9(6):843–851
3. Irsoy O, Cardie C (2014) Opinion mining with deep recurrent neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP-14), pp 720–728
4. Hung BT (2018) Vietnamese keyword extraction using hybrid deep learning methods. In: proceedings of the 5th NAFOSTED conference on information and computer science—NICS
5. Lops P, De Gemmis M, Semeraro G (2011) Content-based recommender systems: State of the art and trends. In: Recommender systems handbook. Springer US, pp 73–105
6. Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowl-Based Syst 46 109–32
7. Isinkaye FO, Folajimi YO, Ojokoh BA (2015) Recommendation systems: principles, methods and evaluation. Egypt Inf J 16(3):261–273
8. Schafer JB et al (2007) Collaborative filtering recommender systems. In: The adaptive web. Springer Berlin Heidelberg, pp 291–324
9. Dong X, Yu L, Wu Z, Sun Y, Yuan L, Zhang F (2017) A Hybrid collaborative filtering model with deep structure for recommender systems. In: AAAI, pp 1309–1315
10. Girase S, Mukhopadhyay D et al (2015) Role of matrix factorization model in collaborative filtering algorithm: a survey

11. Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. In: Proceedings of international conference on learning representations (ICLR-13): workshop track
12. Mikolov T, IlyaSutskever KC, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26:3111–3119
13. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL-14), pp 655–665
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
15. https://www.kaggle.com/snap/amazon-fine-food-reviews

Bui Thanh Hung received his M.S. degree and Ph.D. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2010 and in 2013. He is currently the Director of Data Analytics & Artificial Intelligence Laboratory, Engineering - Technology Faculty, Thu Dau Mot University. His main research interests are Natural Language Processing, Machine Learning, Machine Translation, Text Processing and Data Analytics.

Crowdsourcing Platform for Collecting Cognitive Feedbacks from Users: A Case Study on Movie Recommender System

Luong Vuong Nguyen and Jason J. Jung

Abstract The aim of this research is to present a crowdsourcing-based recommendation platform called *OurMovieSimilarity* (OMS), which can collect and sharecognitive feedbacks from users. In particular, we focus on the user's cognition patterns on the similarity between the two movies. OMS also analyzes the collected data of the user to classify the user group and dynamic changes movie recommendations for each different user. The purpose of this is to make OMS interact intelligently and the data collected faster and more accurately. We received more than a thousand feedbacks from 50 users and did the analyzes this data to group the user. A group of the users can be dynamically changed, with respect to the selection of each user. OMS now still online and collecting data. We have been trying to enrich the cognitive feedback dataset including more than 20,000 feedbacks from 5000 users, so that the recommendation system can make more accurate analysis of user cognitive in choosing the movie similarity.

1 Introduction

The recommendation system is currently popular and applied in many industrial projects, as well as in many different domains [1]. The most common but severe feature of these recommendation systems is to understand what users are thinking and expect. As an example of a store with millions of products on display, the recommendation system relies on the explicit user behaviors (e.g., reviews and ratings) and implicit ones (e.g., clickstreams) in order to provide their customers with the most relevant product.

However, how can we assure whether the recommendations of the system yield the desired results? If the system does not understand what users think about each product, it will never be able to give them the best experience. Therefore, we consider

L. V. Nguyen · J. J. Jung (✉)
Department of Computer Engineering, Chung-Ang University, 84 Heukseok, Seoul, Korea
e-mail: j2jung@gmail.com

L. V. Nguyen
e-mail: nguyenluongvuong@gmail.com

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_9

that understanding the opinions and feedbacks of the users is important and essential so that the system can understand and make suggestions that are accurate and suitable for their needs.

In particular, we focus on the movie domain, and we realized that there are many ways to get feedback from users who have a lot of experience in watching movies. Most of movie recommendation systems (e.g., IMDB, MovieLens, Netflix, and so on) collect information by asking users to rate movies and to write an opinion about personal feelings. However, many users do not want to write their conclusion carefully because it takes their time. Some of the users do not know how to write their opinion clearly and thoroughly.

We want to have a new approach to collect cognitive feedbacks from users by understanding their cognition in selecting similar movies. Then, we can recognize and predict the similar movie selection patterns of the users. Once we have such a large amount of feedback, we can use the statistical analysis methods to understand why users think these two movies are similar? To answer that question, we aim to build a crowd sourcing platform [2] to collect a large number of datasets on user feedback, as mentioned above. We propose a crowd sourcing-based recommendation platform (called OurMovieSimilarity[1]) and the movie data to create sample information for user select have collected automatically by using IMDB open database, or users add themselves to the OMS through personally suggest function.

The problem is that the system has to interact with users efficiently and to be easy to use. We decided that the OMS's layout and the process must be exciting and straightforward [3]. The data collection process described as follows:

- Firstly, the system's new users will choose a movie they have watched and favorited.
- Second, OMS will suggest a list of 5 similar movies (calculated by the system and accompanied by an α parameter that is the user's selection trend, this parameter has initially defaulted 1).
- In the final step, users will select a movie in 5 movies appearance that they think it is similar to the one they chose at first step.

The movie selection will be a loop with variable ω parameters dynamic closer to the user's selection trend. The calculation to suggest a movie and the addition of ω parameters helps users improve the time to choose the movie similar, and they do not have to do it multiple times to find a similar of a movie m_i and a movie m_j. The data collected will be applied in movie similarity recommendation, as we described in Sect. 4.2.

The outline of this chapter includes five sections. Section 1 is for the introduction, and Sect. 2 discusses the related work on crowd sourcing platforms. In Sect. 3, we address a cognitive approach in the recommendation system. We describe experiences on OurMovieSimilarity platform include the overview, functions of OMS, and the data analysis in Sect. 4. Finally, in Sect. 5, we conclude and make some remarks.

[1] http://recsys.cau.ac.kr:8084/ourmoviesimilarity

2 Related Work on Crowdsourcing Platforms

In 2006, in the article "The Rise of Crowdsourcing" which Jeff Howe published in Wired [4], the term *crowdsourcing* appearance the first time, which is a contraction of *crowd* and *outsourcing*. This quote from the article describes simply what crowdsourcing is and how it's made possible by technological advances. The meaning of the word crowdsourcing itself is a portmanteau of crowd and outsourcing [5, 6]. From an online application perspective, online crowdsourcing platforms are increasingly being used to capture ideas from the crowd. Global companies are adopting crowdsourcing ideas to connect with and get feedback from the users. The success of a crowdsourcing platform largely depends on members and their motivation to participate. Motivation determines the quality and quantity of contributions [7].

Crowdsourcing has become a useful tool for understanding audience preferences and anticipating needs [8], which are very important for businesses that depend on innovating or enhancing their products such as fashion brands, food manufacturers, or restaurants [2]. For example, Starbucks provides a simple fill form where people can suggest new ideas, enhance existing services, or request product deliveries. More complex, Amazon MTurk and CrowdFlower [9], which are crowdsourcing platforms, have several jobs such as data categorization, metadata tagging, character recognition, voice-to-text transformation, data entry, email harvesting, sentiment analysis, ad placement on videos, or surveys.

ResearchGate [10] was founded in 2008 by Ijad Madisch, who aims to transform the way researchers are doing their research [11]. For the first time, ResearchGate's headquarters started in Boston and is currently based in Berlin, Germany, and invested by several US venture capital companies. Researchgate now has more than 7 million members., with an average of seven researchers signing up per minute. This system is a crowdsourcing platform for scientists and researchers to ask, answer the question, finding collaborators, and share papers. Although reading an article does not require registration, those who want to become a website member need to have an email address at an accredited or manually verified organization that is a published researcher to register. Each member of the site has a user profile and can upload research results, including papers, data, chapters, adverse effects, patents, research proposals, methods, presentations, and software source code. Besides, users can also follow the activities of other users and participate in discussions with them. Users can also block interactions with other users.

ResearchGate's success has allowed researchers to disseminate their ideas and share their publications for free so they can create collaborations between researchers from around the world. Through this crowdsourcing flatform, members can maintain their publications, ask and answer research-related questions, and follow other researchers to receive their publication updates. This system is capable of tracking the activities of researchers identified as related to each other (such as co-author of the article, project, chapter, and so on.) to send automatic notification emails about updates that associated members. And in fact, this system also has ways to summarize the joint activities of a research group to assign a unique email to limit the number of duplicate emails sent to each researcher.

3 A Cognitive Approach to Recommendation

The first question we asked was, "*why users think two movies are similar?*". Recent advances in psychology and cognitive sciences support the notion that people use a dual-process model, whereby a sense of similarity built on a combination of feature-based taxonomic and relationship-based thematic relations [12]. Taxonomic or hierarchical relationships based on internal characteristics, such as the features of the items themselves, while a thematic relationship is external, which is a separate event or scene that connects the two items. For example, pen and pencil are taxonomically similar since they share many features: both using for writing and under the category of "stationery". Pen and papers are thematically similar since they often used during the same event.

Nowadays, people have witnessed a massive increase in electronic service, e-commerce, and e-business applications operating on the Internet [13]. Recommended systems are increasingly being used by application vendors to make recommendations to their customers [14]. However, most traditional recommendation systems primarily focus on extracting and recommending general priorities based on historical user data [15]. Although users, in general, shared interests can be considered relevant, an individual user also has his personal preferences. He/she can also answer domain expert knowledge to some extent to make a decision. Moreover, very often, while using traditional recommendation systems, users do not readily distinguish whether the items on the page are actual recommendations or merely the content of the page is displayed indiscriminately for all users. Therefore, traditional recommendation systems do not give customers the feeling of being treated individually. Jeff Bezos, CEO of Amazon.com, concluded that if I had 3 million customers on the Web, I should have 3 million stores on the web [15]. Personalized proposal agencies are emerging to overcome the personal nature of integrated recommendations by using technology to assist customers in making decisions about how to treat each customer individually [16].

4 Experiences on OurMovieSimilarity Platform

4.1 Overview

We create OurMovieSimilarity (OMS) platform, which is the crowdsourcing system to collect similar movies from a large amount of users' cognitive data in the selection of similar movies. OMS collects data of movies from the data warehouse provided by IMDB and allows users to have the experience of choosing similar movies from the movies they have seen and also their favorite ones. Figure 1 shows the main features of the OMS system: data collection from the internet and processing of data obtained from users' cognitive of similar movies.

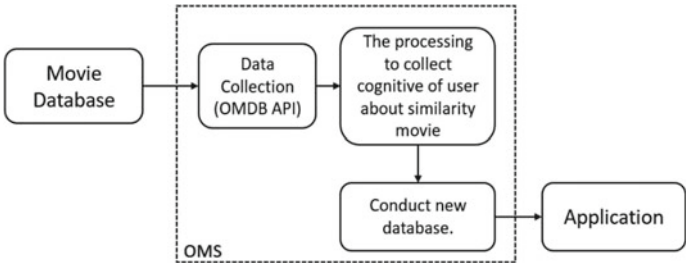


Fig. 1 The overview of the OMS crowdsourcing platform

OMS system is built based on Java programming language, combined with the MySQL database. OMS contains two services, which are web service and background service. The system designed by the Model-View-Controller (MVC) model to ensure security and many flows to handle multiple access and tasks with extensive data. We implement the web service side with Apache Tomcat Servlet Container and Java Server Pages technology. For connecting two services, we choose the Apache Thrift framework. The main features of the OMS system: data collection from the internet and processing of data obtained from users' cognitive of similar movies. Specifically, we have two methods of collecting data about movies, first is to collect automatically from the IMDB database, users add second. Then, OMS collects data from users' selection about the similarity of the movie. Finally, we have a new data set of user perceptions after performing the analysis of the collected data.

The primary purpose of OMS is focussing on simple and efficient. However, OMS is a web-based crowdsourcing platform, and therefore, the problem of loading data from the system to display to users must ensure the lowest latency. Besides, a big problem for a web-based application is making sure that the user has enough instructions to operate the entire system. To solve the issues which we mentioned above, we apply the concept of progressive disclosure to OMS: "show users what they need when they need, and where they want" is the main idea of it. During this period, we focused on increasing the amount of feedback from users most efficiently and accurately. To keep the system running smoothly, and improve the user experience, we take an approach for users to interact simplest and receive the fastest response from the system.

Notably, we apply one template for all pages to maintain the consistency of the user interface. Hence, similar signs will be easily recognized by users (e.g., buttons, functions) while they operate on OMS. We are also applicable this thought to all actions occurring on the system. That is, with all the different steps, we all guarantee the same interface behavior. All of the features we mentioned above based on three gold rules of user interface design [17].

4.2 How Can OMS Interact with Users Intelligently

After users select movies which they have seen and favorite, OMS will calculate and suggest the list of 5 movies similar to their selection. Given m_1 is a movie which users have seen and m_2 is one of all movie remain in our database, we aim to calculate score $Sim(m_1, m_2)$ by making a score after comparing each feature of two movies. The formula for scoring $Sim(m_1, m_2)$ described as follows:

$$Sim(m_1, m_2) \equiv \langle T,\ G, D, A, P \rangle \tag{1}$$

where T, G, D, A, represents the feature of comparing between tile; genre; director; actors; and plot of a movie m_1 and a movie m_2. We repeat that calculation for all the remaining movies in the OMS and obtained a set $\{Sim(m_1, m_2); Sim(m_1, m_3); \ldots; Sim(m_1, m_n)\}$. Then, we add an element that is the trend of users while they select similar movies. Hence, the formula has rewritten as follow:

$$Sim(m_1, m_2) = \frac{\sum_{i=1}^{n} \omega_i \times Sim_i(m_1, m_2)}{\sum_{i=1}^{n} \omega_i} \tag{2}$$

where ω_i are user's trend of selecting similar movies based on i whicha number of features: title, genre, director, actors, and plot, and Sim_i is a score between movie m_1 and movie m_2 based on each feature. To perform the making score of Sim_i after comparing the similarity of each feature, we use a string comparison algorithm base on the Jaccard similarity coefficient (originally coined coefficient by Paul Jaccard) [18]. We do the calculation by identifying the union (characters in at least one of the two sets) of the two sets and intersection (characters which are present in set one which is present in set two). To generalize the calculation formula, we define the following:

$$Sim_X(m_1, m_2) = \frac{\left|X_{m_1} \cap X_{m_2}\right|}{\left|X_{m_1} \cup X_{m_2}\right|} \tag{3}$$

where X is a feature of each movie m_1 and movie m_2. In particular, the score of titles features and described as follows.

$$Sim_T(m_1, m_2) = \frac{\left|T_{m_1} \cap T_{m_2}\right|}{\left|T_{m_1} \cup T_{m_2}\right|} \tag{4}$$

where T_{m_1} represents the title of the movie m_1, and T_{m_2} represent the title of the movie m_2. We perform this calculation for the features: title, genre, director, actor. In case of making a score by comparing the plot of the movie m_1 and movie m_2. Because, it is not a regular string to compare, we pre-processing first. So, we present Term Frequency–Inverse Document Frequency (TF-IDF) to select high weight terms from the plot of

the movie. First, we count the number of times each word occurs in the plot of the movie following the definition of Hans Peter Luhn (1957): "The weight of a term that occurs in a document is simply proportional to the term frequency" [19]. Number of times that term occurs in the plot of movie denoted by raw count. Second, because some term *the; a; an*, and so on are common and tend to incorrectly emphasize scenario which happens to use this term more frequently without giving enough weight to the more exact another time. Hence, following the definition of Karen Sparck Jones: "*The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs*" [20] we use inverse document frequency (IDF) measure how much information the term provides.

Finally, when we use TF-IDF combining with the defined threshold for filtering the useless term ∂, we can find the high weight term frequency in the plot of the movie. So, the formula is shown as follows:

$$P_{m_i} = \left\{ t | \int\left(t, p_{m_i}\right) \times \log \frac{N}{\left|\left\{p_{m_i} \in M : t \in p_{m_i}\right\}\right|} \geq \partial \right\} \tag{5}$$

where P_{m_i} represent all high-frequency term t in a plot of the movie m_i; t is a term that occurs in the plot of a movie m_i; p_{m_i} is all term in the plot of a movie m_i; N is the number of total plots; M is the whole term in all plots remaining; ∂ is the defined threshold for filtering useless term.

By applying TF-IDF, we have obtained for each film a plot of high-meaning terms. So that the comparison between two similar films according to plot feature has more accuracy. Of course, this function wouldbe done entirely automatically, and it means when we have a new movie added to the OMS system, all data, when retrieved, are recalculated to meet the highest accuracy. Beside, OMS will auto dynamic show the suggest movies to users for selection based on the trend of their choice. For example, when user select movie m_1 is "have seen". In the first time, we use our calculating to show movies for users' selection, and the OMS will show a set $\{m_i | i \in [2, .., 6]\}$, and users perhaps have some refresh to select a movie similar. However, after some round limited, we can predict the trend of users' selection described in Fig. 2; and when a user select movie $m_1^{'}$ is "have seen" we will show $\left\{m_i^{'} | i \in [2, \ldots, 6]\right\}$ for a user to select a movie that is more similar than $m_1^{'}$. At this time, showing movies will be dynamically calculated and displayed depending on the users' selection trend. Hence, a user quickly to select similar movie after changing suggested movies one or two times.

Considering the aspects of the relevance of the users in our system, we group related users through the analysis of each user's data. Suppose, users U_i and U_j focus on the priority features to choose the movie similarity in the order <*Title, Genre, Actor, Director, Plot* > ; the system identifies these two users related. Otherwise, the user U_i focuses on <*Title, Genre, Director, Actor, Plot* > and U_j focuses on <*Title, Genre, Plot, Director, Actor* > , the system thinks that these two users are likely to be related together. In this case, the system will calculate and assesses the relevance of these two users for grouping. Therefore, we use soft cosine to measure the similarity between

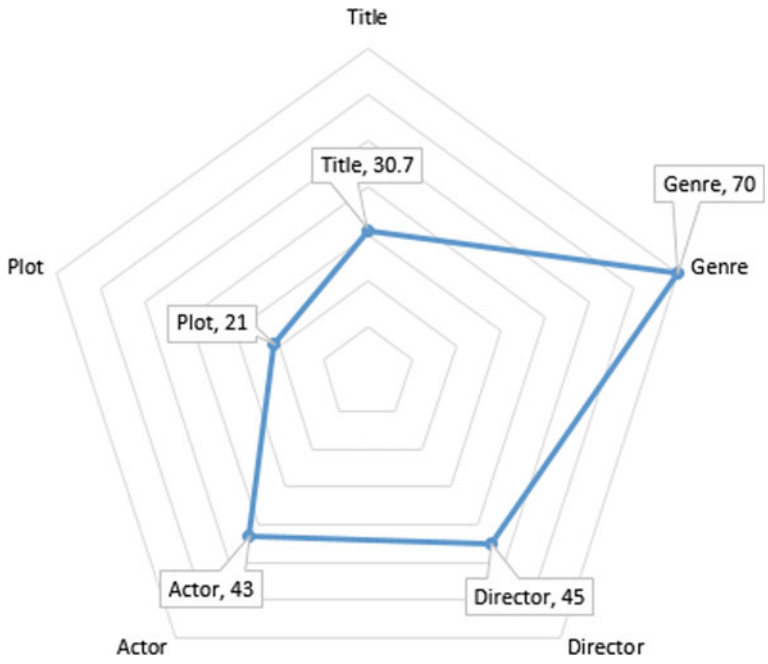


Fig. 2 The trend of the user's selection

two users. The formula to calculate the related between two users as following:

$$Sim(U_1, U_2) = \frac{\sum_{i,j}^{F} s_{ij} U_{1_i} U_{2_j}}{\sqrt{\sum_{i,j}^{F} s_{ij} U_{1_i} U_{1_j}} \sqrt{\sum_{i,j}^{F} s_{ij} U_{2_i} U_{2_j}}} \tag{6}$$

where s_{ij} is the similarity of each feature between two users; F is the number of user's features. From that assumption, we do user grouping and use the data of users in the same group to recommend their chosen movie pairs. For example, user U_i selects the movie (m_i, m_j) the same and user U_i is related to user U_j, the system will suggest that user U_j respond to pair of movie (m_i, m_j) whether or not? The user grouping is shown in Fig. 3.

4.3 Statistical Analysis of the Data (Cognitive Feedbacks)

There are many sources to gather information about movie data provided online. However, we identify an IMDB is an extensive and highly scalable movie database. We collected over 20,000 popular movies from 1990 to 2019 with nine genres and continue importing from open movie database IMDB, 3439 directors, and 8057 actors. The statistics of movie data in OMS is described in Table 1.

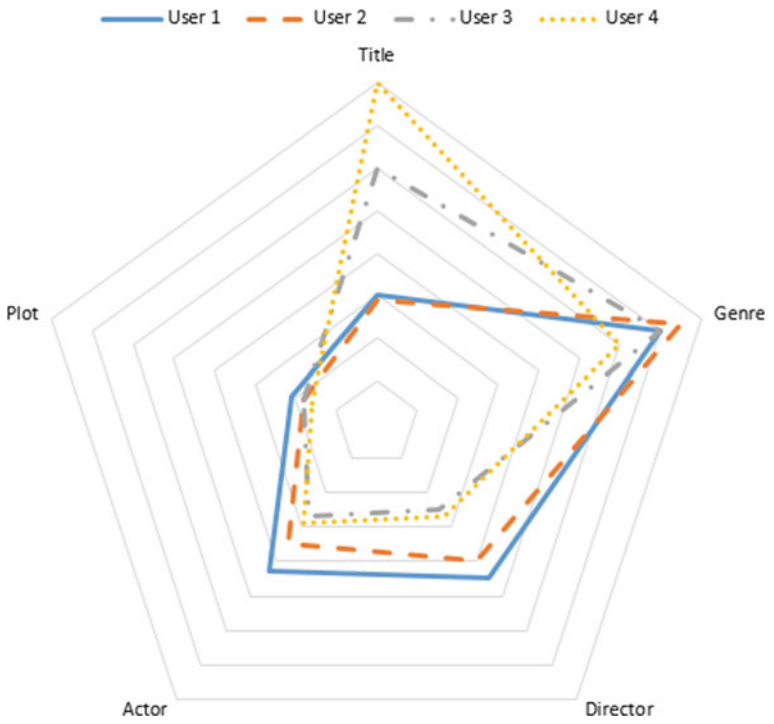


Fig. 3 The related of users based on their trend in selecting movie similarity

Table 1 Statistics of movies data in OMS

#	Title	Amount
1	Number of movies	20,000
2	Range of releases	1990–2019
3	Number of genres	9
4	Range of directors	3439
5	Range of actors	8057

Our system now online and continue collecting data, at this time, we have over 50 users and already raised more than 1000 pairs of movie similarity from users' activities.

The number of data collected from users is shown in Fig. 4, and has the format: $(U_i, m_j, m_k, Sim(m_j, m_k), \vartheta_i)$ inside U_i is the id of the user; m_j and m_k are a pair of the movie similar; $Sim(m_j, m_k)$ is a similarity score we calculated; and ϑ_i is the number of times user change suggested movies in select a pair of movies similar.

We have also done user grouping, but with the number of users in this experiment, the user grouping is too fragmented. We aim to reach a large number of users so that clustering will enhance more efficiency. Figure 5 shows a graph representing several typical user groups in our system.

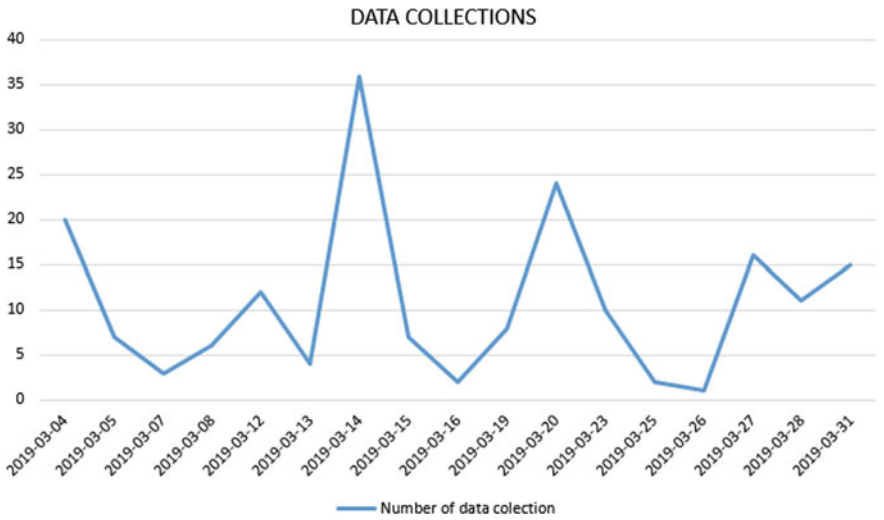


Fig. 4 Statistics amount of data feedback from users by day

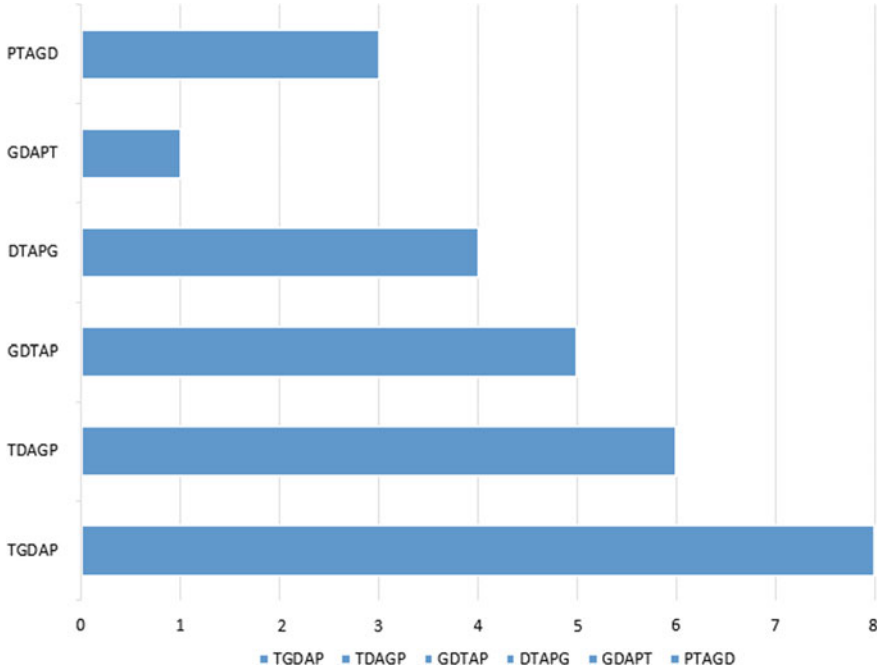


Fig. 5 Clustering of users with T, G, D, A, and P represented sequence title; genre, director; actor; and plot

5 Concluding Remarks

Our study focuses on understanding the cognitive of users and answers the question, "what makes users think two movies are similar?". To do that, in the first stage, we deploy OurMovieSimilarity as a crowdsourcing platform to collect cognitive data from a crowd of users.

Besides, OMS has performed the calculation of data collected by each user to analyze and predict the trend of the user's selection. The data collected shows that the number of page refresh in choosing a pair of movies similarity reduced for each user. This helps to minimize the time, boring, and increase the ability to collect user feedback. From the approach to interacting with such a user, we deduce that people will be motivated to contribute feedback to help others, as well as discover for themselves.

OurMovieSimilarity is still deploying online and continuing to collect data. We plan to have at least 5,000 users and 20,000 responses to make more accurate analyzes of user cognitive in choosing the movie similarity. We release dataset representing user perception of similarity between two movies at the OMS website.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A2B4010774).

References

1. Ricci F, Rokach L, Shapira B (2011) Introduction to recommender systems handbook. In: Recommender systems handbook. Springer, pp 1–35
2. Zheng H, Li D, Hou W (2011) Task design, motivation, and participation in crowdsourcing contests. Int J Electron Commer 15(4):57–88
3. Baturay MH, Birtane M (2013) Responsive web design: a new type of design for web-based instructional content. Procedia-Soc Behav Sci 106:2275–2279
4. Howe J (2006) The Rise of Crowdsourcing. Wired Mag 14(06):1–5
5. Schenk E, Guittard C, Others (2009) Crowdsourcing: what can be outsourced to the crowd, and why. In: Workshop on open source innovation, Strasbourg, France, vol 72, p 3
6. Estellés-Arolas E, González-Ladrón-de-Guevara F (2012) Towards an integrated crowdsourcing definition. J Inf Sci 14
7. Janzik L (2010) Contribution and participation in innovation communities: a classification of incentives and motives. Int J Innov Technol Manag 7(03):247–262
8. Gadiraju U, Kawase R, Dietze S, Demartini G (2015) Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 1631–1640
9. Difallah DE, Catasta M, Demartini G, Ipeirotis PG, Cudré-Mauroux P (2015) The dynamics of micro-task crowdsourcing: The case of amazon mturk. In: Proceedings of the 24th international conference on world wide web, pp 238–247
10. Yu M-C, Wu Y-CJ, Alhalabi W, Kao H-Y, Wu W-H (2016) ResearchGate: An effective altmetric indicator for active researchers? Comput Human Behav 55:1001–1006
11. Dolan KA (2012) How Ijad Madisch aims to disrupt science research with a social network. List Forbes (2012)

12. Wisniewski EJ, Bassok M (1999) What makes a man similar to a tie? stimulus compatibility with comparison and integration. Cogn Psychol 39(3–4):208–238
13. Kalakota R, Robinson M (2001) e-Business 2.0: Roadmap for success. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©2001
14. Ben Schafer JRJ, Konstan J (1999) Recommender systems in E-commerce. In: Proceeding of ACM conference on electronic commerce (EC-99), pp 158–166
15. Ben Schafer J, Konstan JA, Riedl J (2001) E-commerce recommendation applications. Data Min Knowl Discov 5(1–2):115–153
16. Häubl G (2001) Recommending or persuading? the impact of a shopping agent's algorithm on user behavior. In: Proceedings of the 3rd ACM conference on electronic commerce, pp 163–170
17. Mandel T (1997) The elements of user interface design, vol 20. Wiley, New York
18. Bouchard M, Jousselme A-L, Doré P-E (2013) A proof for the positive definiteness of the Jaccard index matrix. Int J Approx Reason 54(5):615–626
19. Luhn HP (1957) A statistical approach to mechanized encoding and searching of literary information. IBM J Res Dev 1(4):309–317
20. Jones KS (2004) A statistical interpretation of term specificity and its application in retrieval. J Doc

Luong Vuong Nguyen is a Ph.D. Student in Chung-Ang University, Korea since August 2018. He received the B.S. in Department of Mathematics from Da Nang University of Education and Science, Vietnam in July 2009. And he received M.S. degrees in Department of Computer Engineering from The Da Nang University of Technology in Vietnam in April 2013. His research topics include knowledge engineering on data science by using data mining, machine learning, ambient intelligence, and logical reasoning.

Jason J. Jung is a Full Professor in Chung-Ang University, Korea, since September 2014. Before joining CAU, he was an Associate Professor in Yeungnam University, Korea since 2007. Also, He was a postdoctoral researcher in INRIA Rhone-Alpes, France in 2006, and a visiting scientist in Fraunhofer Institute (FIRST) in Berlin, Germany in 2004. His research topics are knowledge engineering on social networks by using many types of AI methodologies, e.g., data mining, machine learning, and logical reasoning. Recently, he have been working on intelligent schemes to understand various social dynamics in large scale social media.

A DD-SHELL HF Model for Bus Accidents

Kelvin K. F. Po and Eric T. T. Wong

Abstract Human Factor and Human Error are main reasons that lead to road accidents in Hong Kong. Apart from medicated or drunk cases, human factors are still affecting the road safety in most situations, including driver negligence, environmental influence, etc. As a result, the study of elimination of human error in safety issue became the main study in modern transportation research field. In light of the fact that Human Factor models such as the SHELL and Dirty Dozen have been used successfully in the aviation industry, both of them were applied to analyze a bus accident occurred in February 2018. To clarify the human factor affection and its role in the accident, the technical specification difference between single deck and double-decker buses, as well as other factors like weather and time are neglected so the analysis is solely based on the human factor effect, and the conclusion and recommendation are generated from the investigation result.

Keywords Dirty Dozen · SHELL Human Factor Model

1 Introduction

In most countries, traffic accident takes a top position in the injury and fatality statistics data. In Hong Kong, traffic accidents are usually caused by passenger/driver factor [1], which are defined as human errors. Though Hong Kong possesses a well-developed transportation system, serious accidents persisted in past few years. According to statistics from Hong Kong Transportation Department, driver/passengers factor is a major casualty contributory factor of road accidents.

In the evening of 10 February 2018 a double-decker bus flipped into its side during a downslope turn from a main road. Total fatality was nineteen. In the following day a Commission of Inquiry was set up. The official report [2] made 45 recommendations,

K. K. F. Po · E. T. T. Wong (✉)
Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China
e-mail: mmttwong@connect.polyu.hk

K. K. F. Po
e-mail: kelvin.kf.po@connect.polyu.hk

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_10

covering the training of bus drivers, guidelines on bus driver working hours, rest times and meal breaks, other employment of part-time bus drivers, provision of rest and toilet facilities for bus drivers, abuse and assaults on bus drivers, route risk assessment and speed limits. Based on the world-wide application of aviation human factor models in accident analysis, one can see from the following analysis that there are some gaps to be addressed.

In the following section brief concepts of the SHELL Model and Dirty Dozen Factors are given prior to the analysis.

2 Human Factor Models

2.1 SHELL Model

It is a conceptual framework proposed by the International Civil Aviation Organization [3]. The concept (the name being derived from the initial letters of its components, Software, Hardware, Environment, Liveware) was first developed in 1972.

One practical diagram to illustrate this conceptual model uses blocks to represent the different components of Human Factors in bus operation is shown in Fig. 1, viz.

Fig. 1 The seriously-damaged double decker bus after the crash

- Software—the rules, procedures, written documents etc., which are part of the bus company operating procedures.
- Hardware—the driver's cabins, their configuration, controls, displays and functional systems.
- Environment—the situation in which the bus driver works—the social climate inside bus and the external road environment.
- Liveware—bus driver, passengers, pedestrians, etc.

Of all the dimensions in the model, liveware is the one which is least predictable and most susceptible to the effects of internal (mood, fatigue, motivation, etc.) and external (temperature, light, noise, workload, etc.) changes. Hence Human Error is often seen as the negative consequence of the liveware dimension in this interactive system.

2.2 Dirty Dozen Factors

The Dirty Dozen refers to twelve of the most common human error preconditions, or conditions that can act as precursors, to accidents or incidents. These twelve elements influence people to make mistakes. The Dirty Dozen is a concept developed by Gordon Dupont, in 1993, whilst he was working for Transport Canada, and formed part of an elementary training programme for Human Performance in Maintenance. It has since become a cornerstone of Human Factors in Maintenance training courses worldwide, as exemplified in UK Civil Aviation Authority CAP715 [4].

The original list, developed for aircraft maintenance, is available in many documents, one good example is TC14175 [5], and this list is used as the basis for the current study:

1. Lack of communication—Poor communication often appears at the top of contributing and causal factors in accident reports, and is therefore one of the most critical human factor elements. Communication refers to the transmitter and the receiver, as well as the method of transmission. With verbal communication it is common that only 30% of a message is received and understood.
2. Distraction—Distraction could be anything that draws a person's attention away from the task on which they are employed. Some distractions in the workplace are unavoidable, such as loud noises, requests for assistance or advice, etc.
3. Lack of resources—Resources include personnel, time, data, tools, skill, experience and knowledge etc. A lack of any of these resources can interfere with one's ability to complete a task. It may also be the case that the resources available, including support, are of a low quality or inadequate for the task.
4. Stress—There are many types of stress. Typically in the transport environment there are two distinct types—acute and chronic. Acute stress arises from real-time demands placed on our senses, mental processing and physical body; such as dealing with an emergency, or working under time pressure with inadequate resources. Chronic stress is accumulated and results from long-term demands

placed on the physiology by life's demands, such as family relations, finances, illness, long working hours, inadequate rest, etc. When we suffer stress from these persistent and long-term life events, it can mean our threshold of reaction to demands and pressure at work can be lowered. Thus at work, we may overreact inappropriately, too often and too easily.

5. Complacency—Complacency can be described as a feeling of self-satisfaction accompanied by a loss of awareness of potential dangers. Such a feeling often arises when conducting routine activities that have become habitual and which may be "considered", by an individual (sometimes by the whole organisation), as easy and safe.
6. Lack of teamwork—Many tasks and operations are depending on team efforts.
7. Pressure—Pressure is to be expected when working in a dynamic environment. However, when the pressure to meet a deadline interferes with our ability to complete tasks correctly, then undesirable events may occur. Pressure can be created through lack of resources, especially time; and also from our own inability to cope with an abnormal situation.
8. Lack of awareness—Working in isolation and only considering one's own responsibilities can lead to tunnel vision and a lack of awareness of the effect our actions can have on others. Such lack of awareness may also result from other human factors, such as stress, fatigue, pressure and distraction.
9. Lack of knowledge—lack of on-the-job experience and specific knowledge can lead workers into misjudging situations and making unsafe decisions.
10. Fatigue—We can become fatigued following long periods of work, especially long periods of hard work. As we become more fatigued our ability to concentrate, remember and make decisions reduces. Therefore, we are more easily distracted and we lose situational awareness. Fatigue will also affect a person's mood, often making them more withdrawn, but sometimes more irrational and angry.
11. Lack of assertiveness—Assertiveness is a communication and behavioral style that allows us to express feelings, opinions, concerns, beliefs and needs in a positive and productive manner.
12. Norms—Workplace practices develop over time, through experience, and often under the influence of a specific workplace culture. These practices can be both, good and bad, safe and unsafe; they are referred to as "the way we do things round here" and become Norms. Unfortunately such practices follow unwritten rules or behaviors, which deviate from the required rules, procedures and instructions.

3 Road Traffic Casualties

In the following table, we can observe that the main cause to road accidents in Hong Kong is the driver/passenger factor. From the following Tai Po bus accident [7], it may be seen that both the driver and passengers attributed to the occurrence (Table 1).

Table 1 Road traffic casualities contributory factors

Road traffic casualties by casualty contributory factor and degree of injury 2017				
Casualty contributory factor	Degree of injury			
	Killed	Seriously injured	Slightly injured	Total
Injured by door on bus/PLB	0	0	11	11
Holding vehicle (stealing ride)	1	0	4	5
Fell off platform	0	0	2	2
Opened door negligently (passenger)	0	1	21	22
Closing door negligently (driver/passenger)	0	1	16	17
Passenger lost balance, elsewhere except on stairway of	0	59	516	575
Passenger lost balance, on stairway of bus	0	23	158	181
Lost balance/fell down when boarding/alighting vehicle	0	11	134	145
Other passenger/driver factors	21	463	2,542	3,026
Drunk	0	7	16	23
Under the influence of drugs	0	7	1	3
Suffering from illness, or mental defect	1	1	2	4
Crossing road heedless of traffic (at crossing)	10	40	116	166
Crossing road heedless of traffic (elsewhere)	6	52	198	256
In road, not crossing (jay walking)	0	11	53	64
Listening to audio device	0	0	1	1
Pedestrian inattentiveness	4	78	292	374
Other pedestrian factors	6	26	107	139
No casualty factor	59	1,439	13,376	14,874
Total	108	2,214	17,566	19,888

4 Literature Review on the Application of SHELL Model for Accident Analysis

In December 2014, a study from Holland investigated the human factors affecting the performance of bus drivers, by means of heart rate measurement and questionnaires. Stress, the most influencing factor to the operation condition to bus driver, was being dissolved into several factors to study.

By measuring the stress level for 40-to-60-year-old bus drivers in Stockholm, Prof. Hlotova from KTH in Holland, the researcher of the investigation, discovered that several factors are significant to drivers' mental condition, including weather condition, health status and daily driving perceptions. To explain by Dirty Dozen model, these factors are Pressure, Fatigue, Norm and Complacency respectively. As these factors affects the determination and heart rate which in turn increase the stress level of the driver, they are taken into account during human error analysis and the study was trying to reduce the corresponding human factor influence.

4.1 Investigation Experimental Design

In order to quantify the measurement, Stress level was measured using the HRV (Heart Rate Variability) model, which was a measurement to the change of heart rate with time:

$$HRV = (60/HR)$$

which can directly reflect the interviewee mental status (Table 2).

Table 2 Sample taking and heart rate records [6]

Day	Number of participants	Number of records	Mean	Std. Dev.	Min	Max
Total sample	30	141,163	81.0	11.3	55	138
December	19	93,464	81.6	10.4	58	138
March	11	47,699	79.9	12.8	55	124
4.12.2012, Even-headway	6	21,296	78.9	7.7	58	118
5.12.2012, Snow storm	3	21,191	90.9	8.8	64	138
6.12.2012 December	5	26,145	77.2	10.8	59	124
7.12.2012, December	5	24,832	80.8	7.9	59	116

On the other hand, the sample collected are divided into different control groups, such as smoking and non-smoking, normal and abnormal health status, daily habits, tiredness, etc., which are important parameters to consider the drivers' status in terms of occupational health and condition (Table 3).

Thus, the model analysed by linear regression is conducted and results are shown in Table 4.

The applied least-squares model showed the result of HRV analysis. As deducted from the data, it reflected that drivers' mental condition is largely affected by external environment, e.g. extreme bad weather, which caused a decrease in HRV and hence

Table 3 Variables considered in the data collection [6]

Variable	Mean	Stil. dev
HRV	754.39	102.08
Day with even-headway holding strategy, biliary	0.15	0.36
Day with snowstorm, binary	0.15	0.36
Day after the snowstorm, binary	0.18	0.39
Friday before Christmas, binary	0.18	0.38
Regulation stops, minutes	541.51	502.35
Accumulated time driven, minutes	3899.42	2584.62
Peak Hour, binary	0.24	0.43
Age 24–67 years old. continuous	50.26	8.41
Gender, binary	0.19	0.39
Experience 0–2 years, binary	0.08	0.28
Experience 2–5 years, binary	0.16	0.37
Experience 6–10 years, binary	0.20	0.40
Experience 11–20 years, binary	0.20	0.40
Experience more than 21 years, binary	0.37	0.48
Time within the shift: 0–1 h, binary	0.16	0.37
Time within the shift: 1–3 h, binary	0.44	0.5
Time within the shift: 3–6 h, binary	0.30	0.46
Time within the shift: 6–8 h, binary	0.91	0.29
Medicine influencing heart activity, binary	0.41	0.2
Smoking, binary	0.28	0.45
Everyday coffee, binary	0.78	0.41
Personal feelings of happiness/unhappiness	−0.90	0.87
Personal perception of stressfulness peacefulness	−1.52	0.59
Personal perception of tiredness	−0.48	1.08
Personal perception of driving style	2.45	0.98
Reported time pressure	3.04	0.98

Table 4 HRV model [6]

Variables	Compact model	Shift model	Experience model
Constant	388.11 (307.04)	355.20 (147.80)	219.5 (62.86)
Even-headway strategy	91.96 (2.29)	80.06 (66.57)	147.1 (133.91)
Snowstorm		−62.20 (−65.00)	−28.95 (−22.65)
Day after snowstorm		3.68 (4.17)	−10.03 (−10.23)
Friday, before Christmas	77.19 (1.93)	87.65 (92.58)	150.6 (145.97)
Age	6.74 (4.77)	9.57 (191.94)	9.09 (126.13)
Female		9.00 (8.50)	21.36 (14.94)
Experience: 0–2 years			−31.59 (−16.32)
Experience: 2–5 years			73.68 (74.68)
Experience: 6–10 years			16.61 (14.46)
Experience more than 21 years			35.19(26.86)
Heart-related medicine	195.9 (4.14)	140.90 (125.12)	190.6 (128.35)
Smoking	−99.25 (−3.15)	−55.26 (−89.05)	−108.7 (−114.80)
Coffee every day		−40.51 (−34.46)	−63.35 (−47.52)
Feelings of happiness/unhappiness		9.56 (22.36)	13.36 (13.56)
Perception of stressfulness/peacefulness		60.53 (76.25)	8.718 (12.77)
Perception of tiredness		34.66 (89.09)	25.91 (30.35)
Reported time pressure		0.41 (2.02)	32.17 (88.84)
Perception of driving style			−0.938 (−2.78)
Regulation stops	−0.42 (−18.71)	−0.06 (−3.18)	−0.35 (−17.70)
Accumulated time driven	0.36 (68.43)	0.34 (0.004)	0.35 (78.06)
Peak Hour		−15.21 (−19.80)	−1.778 (2.80)
Elapsed shift: 0–1 h		37.42 (3.40)	
Elapsed shift: 3–6 h		81.21 (112.84)	
Elapsed shift: 6–8 h		68.76 (64.21)	
R-squared	0.376	0.626	0.604
Adjusted R-squared	0.376	0.626	0.604

an increase stress burden, as a typical Liveware-Environment Interface stated in the SHELL model analysis.

Moreover, the study showed that the interaction between drivers' and passengers also increase the stress level, especially when passengers are getting on and off, which is analyzed as Liveware-Liveware interface in the SHELL model, and this experimental result is also applied in the case study in this chapter, since the accident investigated is related to the interaction between the driver and passengers mostly.

4.2 Discussion and Relation to the Study

From the past study, we can see that working environment is significant to the drivers' mental status, namely the Liveware-environment interface of the SHELL model. The influence in external environment, like weather condition, daily life and general driving perception, are causing crucial impact to the drivers' behaviour as it will lead to accidents when we categorize the various sources of stress stated in the Dirty Dozen model. And importantly, the interface between the passengers and the bus driver is also an essential source of stress that lead to a decrease HRV or the Heart rate burden, which in turn leads to a higher probability of human error.

5 Accident Overview

On 10 February 2018, at approximately 18:13 HKT, a Kowloon Motor Bus (KMB) double-decker bus flipped onto its side on Tai Po Road in Tai Po, New Territories. The crash killed 19 people and injured 65. The Volvo Super Olympian bus was running on route 872, a special route operated only on horse racing days. It was travelling from Sha Tin Racecourse, following a race day, to Tai Po New Town.

Before boarding the bus, the driver chatted with a bus fan for some time. It was reported that the bus departed the racecourse 10 min late, leading some annoyed passengers to scold and quarrel with the driver. Passengers said the driver then became frustrated, and "drove really fast as if he was throwing a tantrum". They said he drove "very, very fast" the whole time, without slowing for turns.

As the bus was travelling on a downhill section of Tai Po Road, it came to a bend where it flipped onto its left side, hitting a lamp post and destroying a nearby bus stop. Eighteen people were killed at the scene.

Immediately after the crash, some passengers tried to physically attack the driver. Fire services officers had to cut into the bus to rescue those trapped inside, an operation that took about 85 min.

The day after the crash, the Motor Transport Workers General Union held a press conference to criticize working conditions at KMB. They said that poor pay has led to a shortage of drivers, and alleged that drivers were not provided enough training nor permitted sufficient rest. Similarly, the Confederation of Trade Unions picketed the company's headquarters in protest of alleged low salaries. Chairman of the KMB Staff Union, criticized the company's practice of hiring part-time drivers. He characterized them as amateurish, and said the company only trained them for two days (while full time drivers must complete 18 days of training).

6 Human Factor Analysis

By incorporating the Dirty Dozen into the SHELL model, we have the following analysis of each component:

6.1 Liveware

a. Complacency

Complacency was found on the driver prior to occurrence of the accident. Before the last bus departed from the racecourse, the part-time bus driver was chatting with his colleagues as the driver was complacent to be on time. The delay of departure by 10 min was a leading cause to annoy the passengers, which eventually led to arguments and emotionalized the driver.

b. Pressure

The part-time driver was found stressed under the 10-min delayed schedule. The stress was built up to urge the part-time driver to be distracted by the annoyed passengers under their obscene yelling [5]. Furthermore, quarrel between passenger and the part-time driver sublimed the unstable emotion of the driver.

c. Lack of knowledge

The driver, being a part-time staff, received only 2 days' training instead of 18 days. Hence there is a lack of knowledge on his part.

The passengers onboard, angered by the delayed schedule of bus departure, ignored the published regulation of no disturbance to the driver while the bus is in motion. Such irresponsible act was a cause to distract the driver from driving attentively.

d. Lack of Resources

The working conditions for the bus drivers was blamed by the trade unions.

Poor payment to the full-time drivers attributed to the shortage of drivers. The high turnover of full time drivers has eventually led to the employment of unexperienced part-time drivers. Also, drivers were not given sufficient rest time and meal breaks.

e. Lack of Awareness

Owing to the change of mood resulted from scolding by passengers the driver lost his sense of vigilance. Hence the bus speed was not reduced during a sharp downhill turning and the bus lost control and flipped onto its left side.

f. Distraction of the driver

The no. 872 bus, drove from Shatin Racecourse to Tai Po, was carrying the passengers who may suffer loss in the horsing racing. A noisy and unfriendly atmosphere was thus engendered in the bus. Such a hostile environment would distract and emotionalize the part-time bus driver, who might not be experienced enough to handle the stressful situation.

All of these causes have jointly caused the occurrence of the accident.

6.2 Software

Among other notes, KMB has published the following note to bus passengers:

> Do not speak to the bus captain when the bus is in motion unless it is necessary to do so on the grounds of safety. Do not use any intimidating, insulting or abusive language or act in a manner likely to disturb the bus captain or other passengers on board.

Part-time drivers have not been given adequate training as the full-time drivers.

Current bus schedule does not provide sufficient rest time and meal breaks for drivers.

6.3 Hardware

The design of the driver seat door allows body contact between passengers and the driver, and this would lead to a tensed up situation and hence emotionalise the driver. Angry passengers were not aware that their unruly act would lead to fluctuation of the driver's emotion and the subsequent loss of controllability of the bus during a sharp turn.

Moreover, a lack of speed display in the bus compartment was also crucial to the cause of the disaster. Although many passengers felt that the bus was fast but they were not able to warn the driver based on objective speed data.

6.4 Environment

The noisy and hostile environment inside the bus compartment have caused a distraction to the driver. The speed limit for the road section is 70 km/h.

The resulting DD-SHELL model is shown in Fig. 2. An analysis of the interface between each key components is explained in the following section.

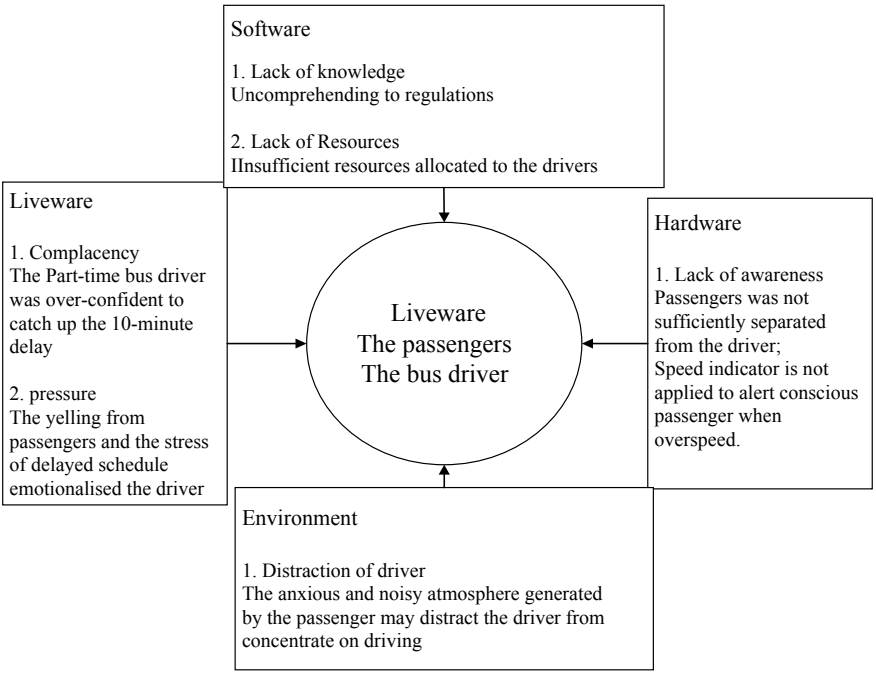


Fig. 2 A DD-SHELL model of bus accident

7 DD-SHELL Interface Analysis

7.1 Liveware-Liveware (L-L Interface)

The part-time driver was over-confident to catch up the schedule, and was chatting with colleagues at the racecourse terminal which led to the stressful situation after the delay. On the other hand, the angered passengers using insulting language towards the part-time driver also deteriorated the situation. The fluctuated emotion of the driver in turn led to the lost of awareness and self-controllability, and after a period of overspeed and overturning, the bus was crashed. With the instinct reflex, the bus driver was turning the bus to the left and led to 19 deaths.

7.2 Liveware-Software (L-S Interface)

The lack of knowledge of the passengers in terms of the ignorance of passenger regulations was an important factor that led to the accident. The passenger neglected the regulations and yelled on the driver during driving was distractive to the driver, who was unaware of the downhill slope condition.

Fig. 3 Double-deck bus driver seat design

Also, lack of resources by the franchised bus company was also a long-term cause of the accident. The underpaid condition of full-time drivers, lack of rest time for drivers and insufficient training provided for part-time driver are latent factors that lead to the vehicle accident. The bus driver, ignored the code of practice to chat with colleagues, was a sparking point of the accident.

7.3 Liveware-Hardware (L-H Interface)

The separation method between the driver and the passengers is a meter-high door installed on the floor, which allowed a direct physical contact between the driver and the passengers. At the time when accident occurred, the driver was arguing with the passengers with unruly languages and acts, which also consisted of bodily contact and this insufficient isolation of driver from passenger needs further attention by the bus company (Fig. 3).

7.4 Liveware-Environment (L-E Interface)

Most of the passengers, being emotionally affected by the result of the last horse race, were angered by the delayed schedule due to the driver's chatting with colleagues. Consequently quarrels were triggered. Under the atmosphere of discontentment, the driver was tensed and lost his vigilance in controlling the bus.

8 Recommendations

Inferred from the above DD-SHELL human factor analysis, a number of recommendations can be made:

(1) The guideline for driver's practice should be reviewed and revised to ensure proper behavior of the driver, no matter full-time or part-time driver. A driver seat recording camera should be installed in all buses to monitor driver for their improper act. On the other hand, it can protect the bus driver from being assaulted by some unruly passengers.
(2) The meter-high door, though convenient for bus drivers and enhanced the accessibility, was not protective and isolative between passengers and the driver. To improve, the franchised bus company should take reference of aircraft cockpit door design, e.g. fully enclosed type door, to ensure a sufficient separation between drivers and passenger (Fig. 4).
(3) An overspeed display and alarm should be installed for passengers, in order to acknowledge passenger at the situation of overspeed, which can serve asawarning to the bus driver and ensure bus controllability (Fig. 5).
(4) An education campaign should be provided to citizens, e.g. TV and media advertisements on passenger regulation and public safety.

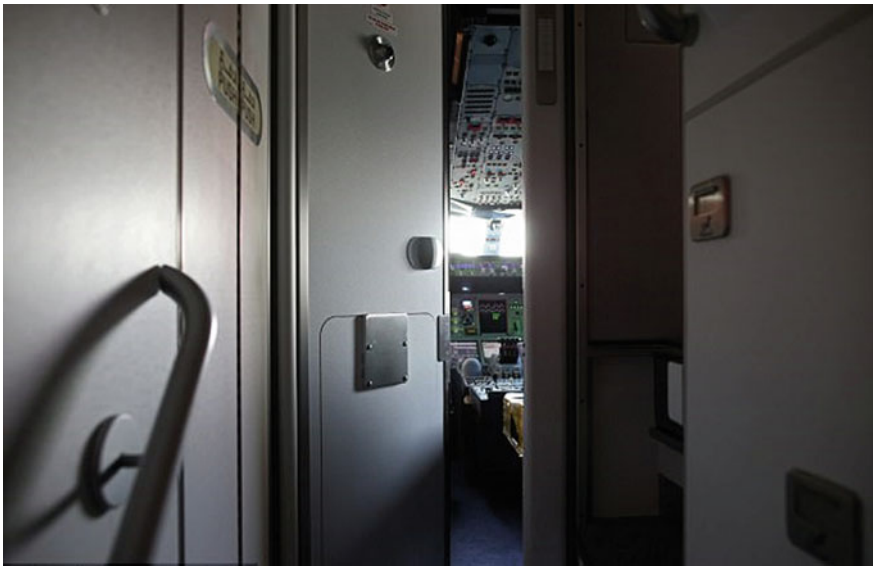


Fig. 4 Aircraft cockpit door

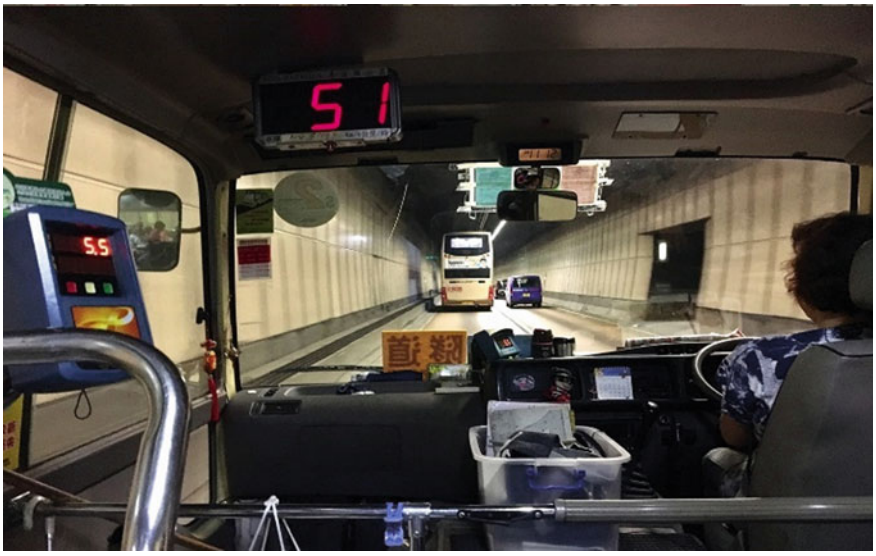


Fig. 5 Hong Kong Minibus speed indicator

9 Conclusion

The disastrous bus accident, occurred on 10 Feb 2018, had led to 19 fatalities and over 60 injuries. The accident was mainly attributed to a combination of human errors. After one year the official report of the Commission of Inquiry published a report on the accident and its contents mainly cover public and organizational issues such the training of bus drivers, bus driver working hours, rest times and meal breaks, other employment of part-time bus drivers, provision of rest and toilet facilities for bus drivers, abuse and assaults on bus drivers, route risk assessment and speed limits. However, the human factor issues are considered incomprehensive. This chapter therefore combines the aviation Dirty Dozen Factor with the SHELL model to analyze the causes of the accident with respect to both human factors and organizational lapses.

Based on our analysis, it is imperative that bus organizations should inculcate a working environment that reiterates these human factors holistically so that they are ingrained in all bus drivers. This can be done through human factor training workshops by reiterating these factors and showing the road accidents they have caused. The aim is to create a safety culture in which all personnel in the transportation system are aware of these factors. This would help reduce the high percentage of driver/passengers casualty involved in traffic accidents.

References

1. https://www.td.gov.hk/en/road_safety/road_traffic_accident_statistics/2017/index.html. 03 Oct 2018
2. HKSAR Government (2018) Report of the independent review committee on Hong Kong's franchised bus service
3. CAP 719 Fundamental Human Factors Concepts. (n.d.). http://publicapps.caa.co.uk/docs/33/CAP719.PDF
4. Human Performance Factors For Elementary Work And Servicing. (N.D.). https://www.tc.gc.ca/media/documents/ca-standards/hpf_elementary.pdf
5. Cheung J (2018) Part-time driver an avid bus fan. http://www.thestandard.com.hk/section-news.php?id=192755&sid=11
6. Hlotova Y, Cats O, Meijer S (2014) Measuring bus drivers' occupational stress under changing working conditions. Transp Res Rec J Transp Res Board 2415(1):13–20. https://doi.org/10.3141/2415-02
7. Grundy T, Lai C 19 killed, over 60 injured in Hong Kong double-decker bus crash. Hong Kong Free Press

Kelvin K. F. Po Kelvin received the bachelor degree of mechanical engineering from Hong Kong Polytechnic University in 2015. He continued his study at Hong Kong Polytechnic University and received the M.Sc. in Mechanical Engineering (Aviation) in 2019. Kelvin is working in engineering fields. He is currently a continuous improvement engineer at Alliance Construction Materials Ltd. He is specialized in production improvement and human factor and error analysis.

Eric T. T. Wong Eric received his M.Sc. degree in Plant Engineering from the Loughborough University (UK) and Ph.D. degree from the Leicester University (UK). Prior to joining the Mechanical Engineering Department of the Hong Kong Polytechnic University, he worked for the Hong Kong Civil Aviation Department in the area of aviation safety. His research interests include: aviation safety management, risk modeling, and quality management. He is a Fellow Member of the Royal Aeronautical Society (UK) and a Founding Member of the Aircraft Engineering Discipline, Hong Kong Institution of Engineers. He has been the Treasurer of the Royal Aeronautical Society (Hong Kong) since its inception in September 1983. He was Vice-Chairman of the HK Institute of Marine Technology in 2002 and Program Chair of the ISSAT International Conference on Modeling of Complex Systems and Environments (2007) at Ho Chi Minh City, Vietnam. He is currently an Adjunct Associate Professor of the Hong Kong Polytechnic University and an Associate Editor of the International Journal of Industrial Engineering—Theory, Application and Practice.

Development of MI-ANFIS-BBO Model for Forecasting Crude Oil Price

Quang Hung Do

Abstract Crude oil price forecasting is an important task in the field of energy research because crude oil is a world's major commodity with a high volatility level. This study proposes the Adaptive Neuro-Fuzzy Inference System (ANFIS) with parameters optimized by Biogeography-Based Optimization (BBO) algorithm and Mutual Information (MI) technique for forecasting crude oil price. The MI is utilized to maximize relevance between inputs and output and minimize the redundancy of the selected inputs. The proposed approach combines the strengths of fuzzy logic, neural network and the heuristic algorithm to detect the trends and patterns in crude oil price data, and thus have been successfully applied to crude oil price forecasting. Other different forecasting methods, including artificial neural network (ANN) model, ANFIS model, and linear regression method are also developed to validate the proposed approach. In order to make the comparisons across different methods, the performance evaluation is based on root mean squared error (RMSE), mean absolute error (MAE), relative absolute error (RAE), root relative squared error (RRSE), and correlation coefficient (R). The performance indexes show that the ANFIS-BBO model achieves lower MAE, RMSE, RAE and RRSE, as well as higher R, indicating that the ANFIS-BBO model is a better method.

Keywords Crude oil price · Forecasting · Adaptive Neuro-Fuzzy Inference System · Biogeography-Based Optimization · Mutual Information

1 Introduction

Crude oil is one of the most strategic commodities and plays an important role in the world economy. Crude oil price can have an effect on several macroeconomics factors including GDP, interest rates, exchange rates. Crude oil prices are determined by various factors and with unexpected trends. Only in one decade from 2003 to 2014, the oil prices have fluctuated from estimated $40/barrel to above $140/barrel and then dropped to below $50. Specifically, in less than six months in 2008, world

Q. H. Do (✉)
Faculty of Information Technology, University of Transport Technology, Hanoi 100000, Vietnam
e-mail: hungdq@utt.edu.vn

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_11

oil prices reached $150/barrel and then fell to under $40/barrel [5, 11]. Unforeseen oil price movements pose a risk for both oil-importing and exporting organizations and countries. Therefore, forecasting crude oil price has been getting much attention from researchers, practitioners and institutions. The accurate crude oil price forecasting can ease the effect of fluctuations on an economy and estimate the upcoming trends of inflation and other economic activities in the future. However, crude oil price forecasting is a challenging task because it is strongly correlated with abnormal, unexpected events and complex factors such as changes in oil inventories, oil production, weather, wars and embargoes [4, 13]. Several approaches have been proposed for the crude oil price forecasts. These methods can be grouped into following categories: (1) methods based on econometric models including time series models, financial models and structural models, and (2) methods bases on computational techniques including artificial neural network, support vector machine, fuzzy logic and expert systems [3].

With the assumption that the characteristics of future crude oil price data are similar to historical and current flow data, forecasting methods based on time series models derived the future price from its own historical data. Several well-known models are vector auto-regression models (VAR) [21], auto-regressive integrated moving average (ARIMA) [36] and semi-parametric approach based on GARCH properties [22]. Financial models obtain the relationship between spot and future prices using hypothesis such as efficiency market hypothesis (EMH) [6]. In structure models, the oil price change is considered as an output of a function with inputs which are the most influential variables. Several common explanatory variables are OPEC behavior, oil inventory level, oil consumption and production, exchange rate. The above mentioned models have the ability of providing good forecasting results when the crude oil price data series under study is linear or near linear. However, real-world crude oil price data is with a strong nonlinear and chao time series, there is a great deal of nonlinearity and irregularity. Therefore, the econometrics model seems to be inappropriate for building the forecasting model of crude oil price.

As for the use of computational models, several approaches have been applied to the task of crude oil forecasting. Artificial neural network (ANN) is certainly the most widely used one for forecasting of crude oil price data. The forecasting obtained results by ANN are than those of ARIMA and GARCH [23]. It is a supervised learning algorithm that can be trained to learn a function between input features and the output, which is represented by the target to be predicted. The most widely used ANN-based model in crude oil price forecasting is Feed Forward Neural Network (FFNN). The state of art of using ANN in forecasting crude oil price is discussed and presented in Bashiri Behmiri and Pires Manso's work [3].

Another development of artificial intelligence based techniques is the combination of the ANN and other methods. Shabri and Samsudin [27] combined the strengths of discrete wavelet transform and ANN for daily crude oil price forecasting. In West Texas Intermediate (WTI) and Brent crude oil spot prices, it was found that wavelet ANN model provided more accurate crude oil prices forecasts than individual ANN

model. Although various methodologies have been applied to the oil price forecasting problem, the ultimate objective remains the same: to obtain the forecasting result with high accuracy and robustness. Attention has also focused on improving the existing methodologies and models. The combination of different techniques can maximize the effectiveness and minimize the potential side effects of each one.

In forecasting research, the study on how to select feature get a lot of attention. A feature selection is a procedure of choosing the most effective features from a set of original features. The mutual information has been adopted for selecting features [14]. Mutual Information (MI) theory has the ability of decreasing prediction error covariance by decreasing the dependence on observation error. Like the cross-correlation method used to calculate the non-linear correlation between two quantities, MI estimates the non-linear correlation between two non-linear quantities [18, 37]. The mutual information feature selection (MIFS) algorithm estimates the MI between the input data and the most important data selected for forecasting to decrease the computational complexity and increase the forecasting accuracy. Therefore, MIFS is a suitable method for forecasting crude oil price.

It has been shown that ANNs are more effective than traditional methods in various application areas [9, 10, 26]. Other than that, the hybrid intelligent system, which is based on the combination of artificial neural networks and other intelligent techniques, has been proposed to take the full advantages of ANN. Fuzzy systems are appropriate if sufficient expert knowledge about the process is available, while neural networks are useful if sufficient process data is available or measurable. The ANFIS, a hybrid intelligent system, is a combination of artificial neural networks and fuzzy systems; therefore, it has the advantages of both methods [2, 7, 19]. Fuzzy systems are appropriate if sufficient expert knowledge about the process is available, while neural networks are useful if sufficient process data is available or measurable. ANFIS can effectively solve non-linear problems and is particularly useful in applications where classical approaches fail or are too complicated to be used [7] and is particularly useful in applications where traditional approaches fail or are too complicated to be used [17].

On the other hand, new evolutionary algorithms, including Biogeography-Based Optimization (BBO), inspired by the behavior of natural phenomena, were developed for solving optimization problems. Through the competitive results of benchmarking studies, these algorithms have been proven to be powerful and are considered to outperform other well-known algorithms. The BBO, proposed by Simon [28], was inspired by the migration process of species. Since then, BBO has been used in solving various complicated problems and is considered to outperform other algorithms, such as genetic algorithms (GA), ant colony optimization algorithms (ACO) [12, 20, 34].

The merit of BBO algorithm, the fuzzy system and the success of ANN have encouraged us to combine these techniques for forecasting crude oil price.

2 Preliminaries

2.1 Feature Selection Based on Mutual Information

Feature selection is an initial stage used in machine learning when selecting a subset of features from data. This is a necessary step to avoid the curse of dimensionality. A subset is considered as the best subset if it contains the least number of features providing the accuracy. Mutual information (MI) is one of the most powerful feature selection techniques. The idea of mutual information is based on the concept of entropy, which is an estimate of the uncertainty of random variables.

For discrete random variables X with values X_1, X_2, ..., X_n and probabilities $P(X_1)$, $P(X_2)$, ..., $P(X_n)$, respectively, the entropy H(X) is calculated as follows:

$$H(X) = -\sum_{i=1}^{n} P(X_i) \log_2(P(X_i)) \tag{1}$$

The joint entropy H(X, Y) of two discrete random variables X and Y with a joint probability distribution P(X, Y) is defined by

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P\left(X_i, Y_j\right) \log_2\left(P\left(X_i, Y_j\right)\right) \tag{2}$$

where H(X, Y) is the total of the entropy of random variables X and Y. When the variable Y is known and X is not, the remaining uncertainty is calculated as follows:

$$H\left(\frac{Y}{X}\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P\left(X_i, Y_j\right) \log_2\left(P\left(\frac{Y_j}{X_i}\right)\right) \tag{3}$$

$H\left(\frac{Y}{X}\right)$ is called the conditional entropy. The relation between the joint entropy and the conditional entropy is as follows:

$$H(X, Y) = H(X) + H\left(\frac{Y}{X}\right) = H(Y) + H\left(\frac{X}{Y}\right) \tag{3}$$

This indicates that the total entropy of random variables X and Y is the entropy of X plus the remaining entropy of Y for a given X. The Mutual Information between X and Y can be calculated as follows:

$$MI(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P\left(X_i, Y_j\right) \log_2\left(\frac{P\left(X_i, Y_j\right)}{P(X_i) P\left(Y_j\right)}\right), \tag{4}$$

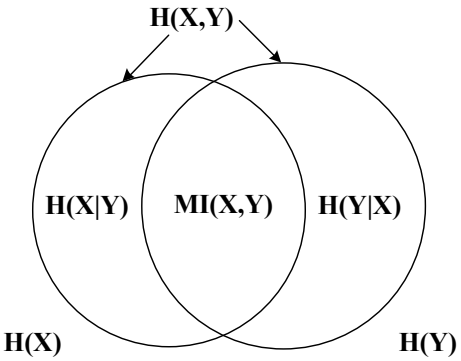


Fig. 1 Representation of mutual information and entropy

where P(X) and P(Y) denote the marginal density functions, P(X, Y) is the joint probability function.

The common information between two random variables X and Y is defined as the mutual information MI(X, Y). Intuitively, MI(X, Y) measures the information that X and Y share: it measures how much knowing one of these variables reduces uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. At the other extreme, if X is a deterministic function of Y and Y is a deterministic function of X then all information conveyed by X is shared with Y knowing X determines the value of Y and vice versa. Mutual information is a values in between 0 and 1, where the larger the value, the greater the relationship between the two variables. If the MI is zero, then the variables are independent (Fig. 1).

2.2 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Neural networks and fuzzy set theory, which are soft computing techniques, are tools for establishing intelligent systems. A fuzzy inference system (FIS) employs fuzzy if-then rules when acquiring knowledge from human experts to deal with imprecise and vague problems [38]. FISs have been widely used in many applications including optimization, control, and system identification [16, 30]. A simple FIS is presented in Fig. 2. However, fuzzy systems cannot learn from or adjust themselves [1]. A neural network has the capacity to learn from its environment, self-organize, and adapt in an interactive way. For these reasons, a neuron-fuzzy system, which is the combination of a fuzzy inference system and neuron network, has been introduced to produce a complete fuzzy rule base system [25]. The advantage of neural networks and fuzzy systems can be integrated in a neuron-fuzzy approach. Fundamentally, a neuron-fuzzy system is a fuzzy network that has its function as a fuzzy inference system. The system can overcome some limitations of neural networks, as well as the limits of fuzzy systems [24], when it has the capacity to represent knowledge

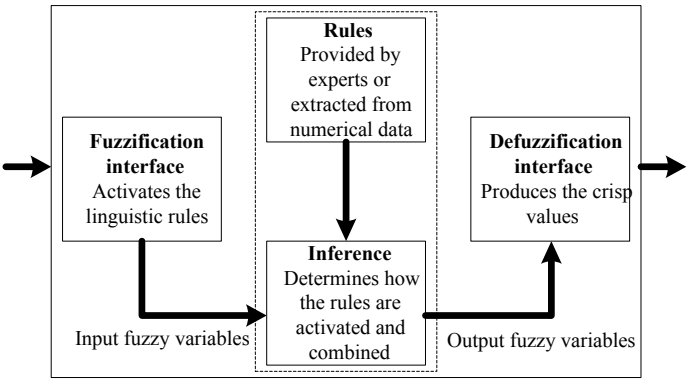


Fig. 2 A simple fuzzy inference system

in an interpretable manner and the ability to learn. The details of the neuron-fuzzy system were proposed by Takagi and Hayashi [32]. Among the neuron-fuzzy systems, ANFIS, proposed by Jang, has been one of the most common tools. In the FIS, the fuzzy if-then rules are determined by experts, whereas in the ANFIS, it automatically produces adequate rules with respect to input and output data, and facilitates the learning capabilities of neural networks.

ANFIS is a multilayered feed-forward neural network, which employs neural network learning algorithms and fuzzy reasoning to map from input space to output space. The architecture of ANFIS includes five layers, namely, the fuzzification layer, the rule layer, the normalization layer, the defuzzification layer, and a single summation node. To present the ANFIS architecture and simplify the explanations, we assume that the ANFIS has two inputs, x and y, two rules, and one output, f, as shown in Fig. 3. Each node within the same layer performs the same function. The circles are used to indicate fixed nodes, while the squares are used to denote adaptive nodes. A FIS has two inputs and two fuzzy if-then rules [33] may be expressed as:

$$\text{Rule 1}: \quad \textit{If } x \textit{ is } A_1 \textit{ and } y \textit{ is } B_1 \textit{ then } f_1 = p_1 x + q_1 y + r_1$$

$$\text{Rule 2}: \quad \textit{If } x \textit{ is } A_2 \textit{ and } y \textit{ is } B_2 \textit{ then } f_2 = p_2 x + q_2 y + r_2 \tag{5}$$

where x and y are the inputs; A_1, A_2, B_1, B_2 are the linguistic labels; p_i, q_i, and r_i, (i = 1 or 2) are the consequent parameters that are identified in the training process; and f_1 and f_2 are the outputs within the fuzzy region. Equation 5 represents the first type of fuzzy if-then rules, in which the output part is linear. The output part may also be constants [31], represented as:

$$\text{Rule 1}: \quad \textit{If } x \textit{ is } A_1 \textit{ and } y \textit{ is } B_1 \textit{ then } f_1 = C_1$$

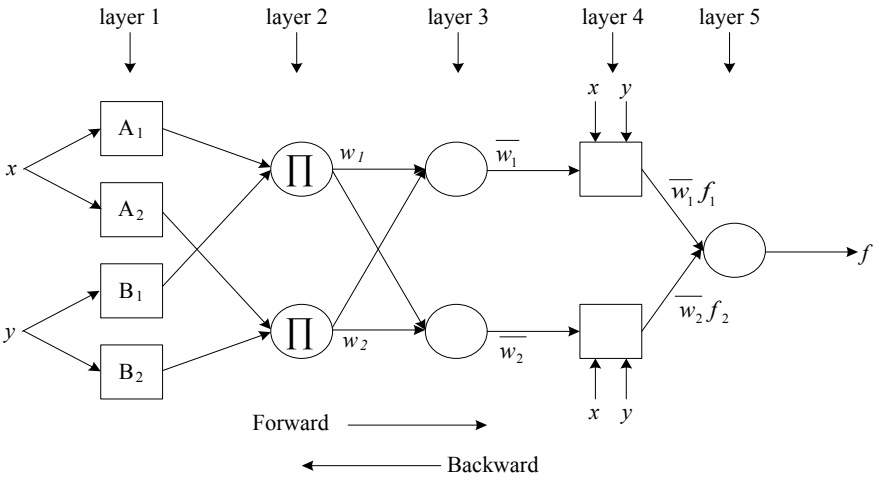


Fig. 3 An ANFIS architecture of two inputs and two rules

$$\text{Rule 2}: \quad \textit{If } x \textit{ is } A_2 \textit{ and } y \textit{ is } B_2 \textit{ then } f_2 = C_2 \tag{6}$$

where C_i ($i = 1$ or 2) are constant values. Equation 6 represents the second type of fuzzy if-then rules.

For complicated problems, the first type of if-then rules is widely utilized to model the relationships of inputs and outputs [35]. In this research, a linear function was used for the output.

The functions of each layer in Fig. 2 are briefly described as follows:

Layer 1—fuzzification layer: Every node in this layer is a square node. The nodes produce the membership values. Outputs obtained from these nodes are calculated as follows:

$$O_{1,i} = \mu A_i(x), \textit{ for } i = 1,\ 2$$

and

$$O_{1,i} = \mu B_{i-2}(y), \textit{ for } i = 3,\ 4 \tag{7}$$

where $O_{1,i}$ denotes the output of node i in layer 1, and μAi(x) and μBi-2(y) are the fuzzy membership functions of A_i and $B_{i\text{-}2}$. The fuzzy membership functions can be in any form, such as triangular, trapezoidal, Gaussian functions.

Layer 2—rule layer: Every node in this layer is a circular node. The output is the product of all incoming inputs.

$$O_{2,i} = w_i = \mu A_i(x) \times \mu B_i(y), \textit{ for } i = 1,\ 2 \tag{8}$$

where $O_{2,i}$ denotes the output of node i in layer 2 and w_i represents the firing strength of a rule.

Layer 3—normalization: Every node in this layer is a circular node. Outputs of this layer are called normalized firing strengths. The output of the *i*th node in this layer is calculated by the *i*th node firing strength to the sum of all rules' firing strengths:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, for\ i = 1,\ 2 \tag{9}$$

where O3,i denotes the output of node i in layer 3, and $\overline{w_i}$ is the normalized firing strength.

Layer 4—defuzzification layer: Every node in this layer is an adaptive node with a node function:

$$O_{4,i} = \overline{w_i} f_i,\ for\ i = 1,\ 2 \tag{10}$$

where $O_{4,i}$ denotes the output of node i in layer 4, $\overline{w_i}$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer are called consequent parameters of the Sugeno fuzzy model.

Layer 5—a single summation node: The node is a fixed node. This node computes the overall output by summing all of the incoming signals from the previous layer:

$$O_{5,i} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i},\ for\ i = 1,\ 2 \tag{11}$$

where $O_{5,i}$ denotes the output of node *i* in layer 5. The results are then defuzzified using a weighted average procedure.

It can be seen that the ANFIS architecture has two adaptive layers: layer 1 and layer 4. Layer 1 has parameters related to the fuzzy membership functions, and layer 4 has parameters $\{p_i, q_i, r_i\}$ related to the polynomial. The aim of the hybrid learning algorithm in the FFN architecture is to adjust all of these parameters in order to make the output match the training data. Adjusting the parameters includes two steps. In the forward pass of the learning algorithm, the premise parameters are fixed, functional signals go forward until layer 4, and the consequent parameters are identified by the least squares method to minimize the measured error. In the backward pass, the consequent parameters are fixed, the error signals go backward, and the premise parameters are updated by the gradient descent method [15]. In the study, the BBO algorithm is utilized in optimizing parameters of ANFIS.

2.3 Biogeography-Based Optimization (BBO) Algorithm

Biogeography is the science which studies the geographical distribution of living species. BBO is a new inspired algorithm that is based on biogeography [28, 29]. Simon [28] developed the mathematical models of biogeography to solve optimization problems. In BBO, variables that determine the quality of habitat are called suitability index variables (SIVs), and each habitat is considered as an individual and has its habitat suitability index (HSI). SIVs are independent variables and HSI depends on SIVs. Habitats with large values of HSI accommodate more species which are suitable for species living, and, conversely, a low-HSI habitat supports fewer species which are not suitable for species living. When the number of species in a habitat increases there is a strong tendency for species to emigrate from the crowded habitat to find new ones with better life-supporting conditions and lower population density than the old habitats. Habitats with low population density may accept a lot of new species from high-HIS habitats by providing adequate life-supporting characteristics. The objective function can be considered as HSI and the evolutionary procedure of BBO is to acquire the solutions which maximize the HSI by using the immigration and emigration features of the habitats. The pseudo code of BBO algorithm can be described in Fig. 4.

```
Initialize the BBO parameters
Generate a set of habitats, corresponding to the potential
solutions
Evaluate the fitness value or HSI for each habitat
While Stopping criterion is not satisfied do
    Decide immigration rate λ and emigration rate μ for each
habitat
    Modify habitats based on λ and μ
     For i=1 to N (population size) do
       Use λ to probabilistically decide whether to modify a
habitat
       If rand(0,1) < λi
         Select habitat Hj to immigration
         Perform migration on Hi and Hj
         Evaluate the fitness value or HSI for newly generated
solution
         Replace the new solution with Hi
       End
       If rand(0,1) < PMutation
         Apply mutation on Hi
         Evaluate the fitness value or HSI for newly generated
solution
       End
     End
    Update habitats' population
End
```

Fig. 4 Pseudo code of BBO algorithm

In BBO, the probability to choose the solution H_i as the immigrating habitat depends on its immigration rate λ_i and solution H_j depends on its emigration rate μ_j. Migration can be demonstrated as:

$$H_i(SIV) \leftarrow H_j(SIV)$$

The immigration rate and emigration rate can be described by Eqs. 12–13.

$$\lambda_i = I\left(1 - \frac{k_i}{n}\right) \tag{12}$$

$$\mu_i = E\left(\frac{k_i}{n}\right) \tag{13}$$

where I and E are the maximum possible immigration rate and emigration rate, respectively. k_i represents the rank of habitat i after sorting all habitats according to their HSI; and n is the number of solutions in the population. A better solution has higher emigration and lower immigration rate, and vice versa.

3 Methodology

3.1 Selecting Features

In feature selection, it is expected to maximize relevance between inputs and output, as well as minimize the redundancy of selected inputs. Crude oil price traded on NYMEX can be considered as non-linear maps about their historical values and several exogenous variables, including gold price, US government treasury bill index, and S&P 500 index. Other than that, several previous studies have showed the impacts of political events of OPEC countries on the crude oil price. As of May 2018, OPEC has 14 member countries: six in the Middle East, six in Africa, and two in South America. According to the U.S. Energy Information Administration (EIA), OPEC accounts for an estimated 44% of global oil production and 73% of the world's oil reserves, giving OPEC a major influence on global oil prices.

The International Country Risk Guide (ICRG) index released by the Political Risk Service (PRS) group has been commonly used as a proxy for the countries' political risk situation [8]. ICRG index includes twelve indicators and each indicator is within between 0 and a maximum value.

The Volatility Index (VIX) is also utilized in oil price forecasting. It is an index created by the Chicago Board Options Exchange (CBOE), which shows the market's expectation of 30-day volatility.

A candidate feature set can be formed as follows:

$$S(t) = \left\{ \begin{array}{r} CO(t-1), CO(t-2), \ldots, CO\big(t-N_{CO}\big), G(t-1), G(t-2), \ldots, G\big(t-N_G\big), \\ TB(t-1), TB(t-2), \ldots, TB(t-N_{TB}), SP(t-1), SP(t-2), \ldots, SP\big(t-N_{SP}\big), \\ ICRG(t-1),\ ICRG(t-2), \ldots, ICRG(t-N_{ICRG}), VIX\,(t-1), VIX\,(t-2), \ldots, VIX\,(t-N_{VIX}) \end{array} \right\}$$

where $CO(t-i)$, $G(t-i)$, $TB(t-i)$, $SP(t-1)$, $ICRG(t-i)$ and $VIX(t-i)$ denote the value of crude oil price, gold price, treasury bond, S&P index, ICRG index and VIX index at $(t-i)$, respectively.

3.2 The Proposed ANFIS with Parameters Optimized by BBO

It can be seen that the ANFIS architecture has two adaptive layers: layer 1 and layer 4. Layer 1 has parameters related to the fuzzy membership functions, and layer 4 has parameters $\{p_i, q_i, r_i\}$ related to the polynomial. The aim of the hybrid learning algorithm in the ANFIS architecture is to adjust all of these parameters in order to make the output match the training data. Adjusting the parameters includes two steps. In the forward pass of the learning algorithm, the premise parameters are fixed, functional signals go forward until layer 4, and the consequent parameters are identified by the least squares method to minimize the measured error. In the backward pass, the consequent parameters are fixed, the error signals go backward, and the premise parameters are updated by the gradient descent method [15]. In the study, the BBO algorithm was utilized in optimizing parameters of ANFIS.

The parameters in the ANFIS that need to be updated is $\Theta = \{p_i, q_i, r_i\}$.

The ANFIS parameters are updated according to the performance index of root mean square error (RMSE) given by:

$$RMSE(\Theta, \tilde{x}) = \sqrt{\frac{1}{K}\sum_{i=0}^{K}\left(y_i - \hat{y}_i\right)^2} \tag{14}$$

where y_i is the actual (desired) value and $\hat{y}_i$ represents the forecasted value.

In the study, the BBO algorithm is utilized to train the forecasting model. The best parameters $\Theta = \{p_i, q_i, r_i\}$ are selected based on performance criteria. First, the whole data set is split into the training set and the testing set. After the training process (as shown in Fig. 5), the trained ANFIS based on the training set is applied to the testing set, and the performance criteria are recorded.

To examine the performance of the proposed model, several criteria are used. These criteria are applied to the developed model to know how well it works. The criteria were used to compare predicted values and actual values. They are as follows:

Correlation coefficient (R): This criterion reveals the strength of relationships between actual values and predicted values. The correlation coefficient has a range from 0 to 1, and a model with a higher R means it has better performance.

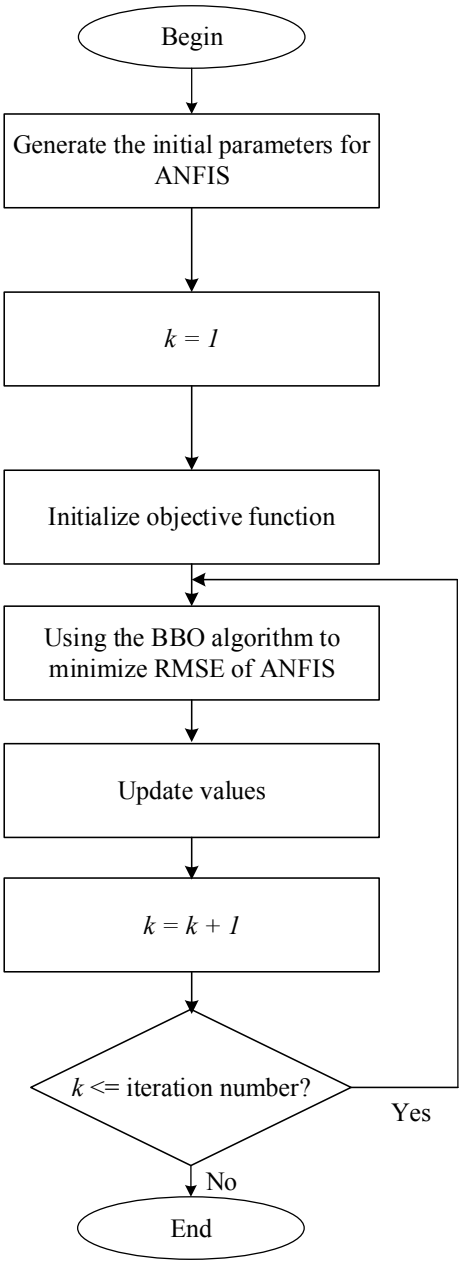


Fig. 5 Training ANFIS by the BBO algorithm

$$R = \frac{\sum_{k=1}^{m} \left(t_k - \bar{t}\right)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m} \left(t_k - \bar{t}\right)^2 . \sum_{k=1}^{m} (y_k - \bar{y})^2}} \tag{15}$$

where $\bar{t}$ and $\bar{y}$ are the average values of tk and yk, respectively.

Mean absolute error (MAE): This index indicates how close predicted values are to the actual values.

$$MAE = \frac{1}{m} \sum_{k=1}^{m} |t_k - y_k| \tag{16}$$

Root mean squared error (RMSE): This index estimates the residual between the actual value and desired value. A model has better performance if it has a smaller RMSE. An RMSE equal to zero represents a perfect fit.

$$RMSE = \sqrt{\frac{1}{m} \sum_{k=1}^{m} (t_k - y_k)^2} \tag{17}$$

Relative absolute error (RAE): This index can be defined as an absolute error (taking the absolute value of residual) as a fraction of the actual value of the outcome target variable.

$$RAE = \frac{\sum_{k=1}^{m} |t_k - y_k|}{\sum_{k-1}^{m} |y_k - \bar{y}|} \tag{18}$$

Root relative squared error (RRSE): The relative absolute error (and analogically RRSE) is calculated as the mean absolute error divided by the absolute error.

$$RRSE = \frac{\sum_{k=1}^{m} (t_k - y_k)^2}{\sum_{k-1}^{m} (y_k - \bar{y})^2} \tag{19}$$

where t_k is the actual (desired) value, y_k is the predicted value produced by the model, and m is the total number of observations.

4 A Case Application

There exists several common crude oil price data. In order to investigate the applicability of the proposed approach, we use the two crude oil price data series including West Texas Intermediate (WTI), and Brent Blend. We collected weekly data series from January 2000 to September 2018, with a total of 970 samples. For example, the data series of the last year (from September 2017 to September 2018) are shown

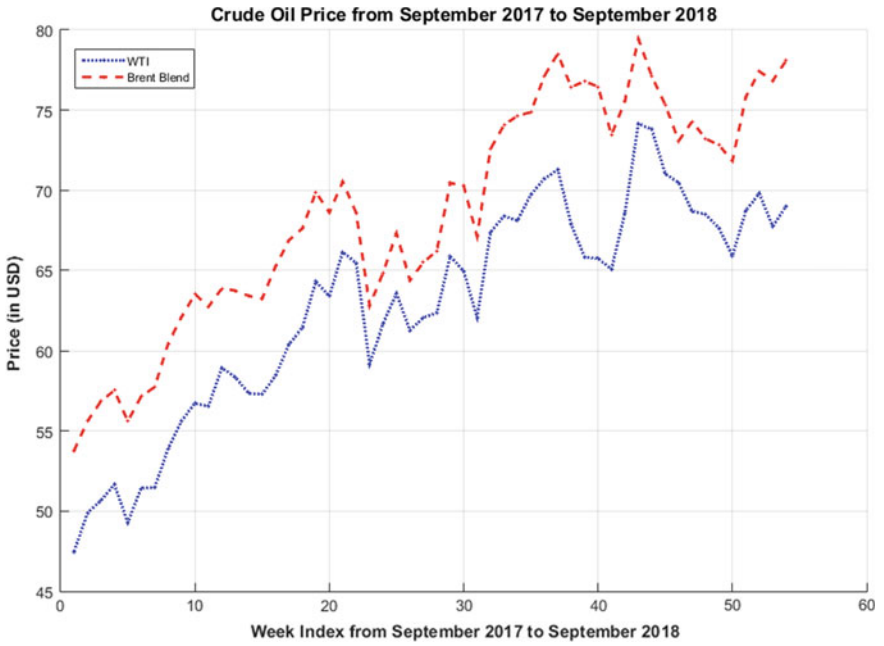


Fig. 6 Weekly crude oil price data from September 2017 to September 2018

in Fig. 5. The weekly oil prices are computed by conducting an average of the daily closing spot prices for a specified product over the given time horizon. The selection of these crude oil price data has been made since their availability and common usage of oil price trends and forecasting. The source of data is the International Monetary Fund (IMF) primary commodity price database, Energy Information Administration (EIA), World Bank, International Energy Agency (IEA), World Energy. This data will then be preprocessed and normalized in order to remove noise and missing values. All the data have been transformed to the 2018 prices in order to remove the differences resulting from inflation (Fig. 6).

5 Results and Discussion

In this study, different forecasting models and the proposed ANFIS-BBO forecasting model are developed and investigated. For each model, we conducted 10 independent tests, and each test produced a set of performance criteria values. The average performance criteria for each model were calculated and are presented in Table 1. The scattering diagrams and price graphics are also drawn in Figs. 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16. In addition to the proposed ANFIS-BBO forecasting model, other forecasting models developed in this study include ANN-based model, ANFIS-based

Table 1 Performance statistics of different forecasting models

Item	Model	R	MAE	RMSE	RAE (%)	RRSE (%)
Brent	ANFIS model	0.9554	7.3514	9.3435	28.134	30.1502
	ANN model	0.9526	7.4522	9.42	28.52	30.3971
	ANFIS-BBO model	0.9865	3.4741	5.0983	13.2955	16.4517
	Linear regression	0.9026	11.2879	13.3283	43.1992	43.0088
WTI	ANFIS model	0.943	7.4832	9.3021	32.5471	34.3668
	ANN model	0.9467	6.7941	8.7128	29.5503	32.1898
	ANFIS-BBO model	0.9828	3.4626	5.0095	15.0601	18.5077
	Linear regression	0.8887	10.3012	12.3977	44.8037	45.8034

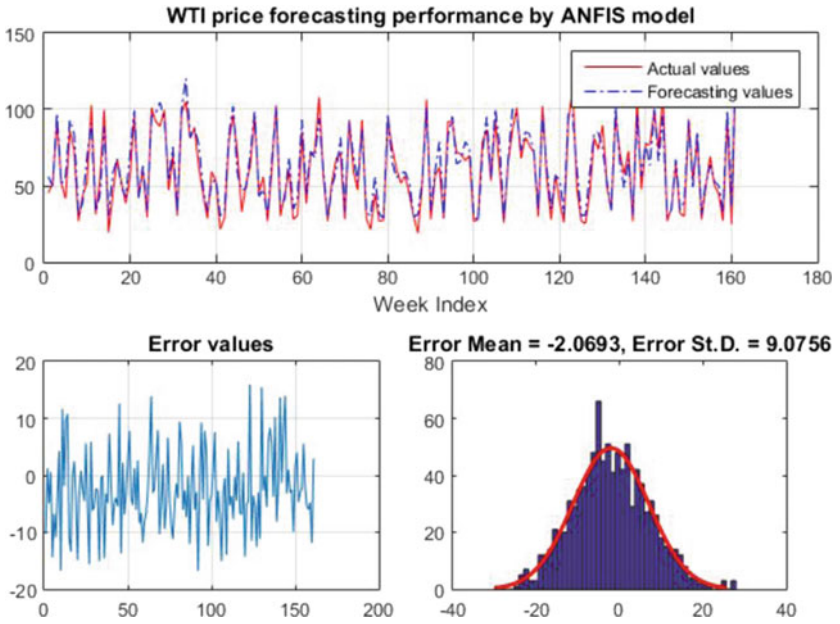


Fig. 7 WTI price forecasting performance by ANFIS model

model, and linear regression method. A 10-fold cross validation method was used to avoid an over-fitting problem.

For ANN-based model, we adopt a feed-forward network (FFN) with one hidden layer to forecast traffic volume. The optimum number of neurons in the hidden layer was determined by varying their number, starting with a minimum of one, and then increasing in steps by adding one neuron each time. Hence, various FFN architectures were tested to achieve the optimum number of hidden neurons. The best performing architectures for ANN were then found. The activation function from input layer to hidden layer is sigmoid. With no loss of generality, a commonly used

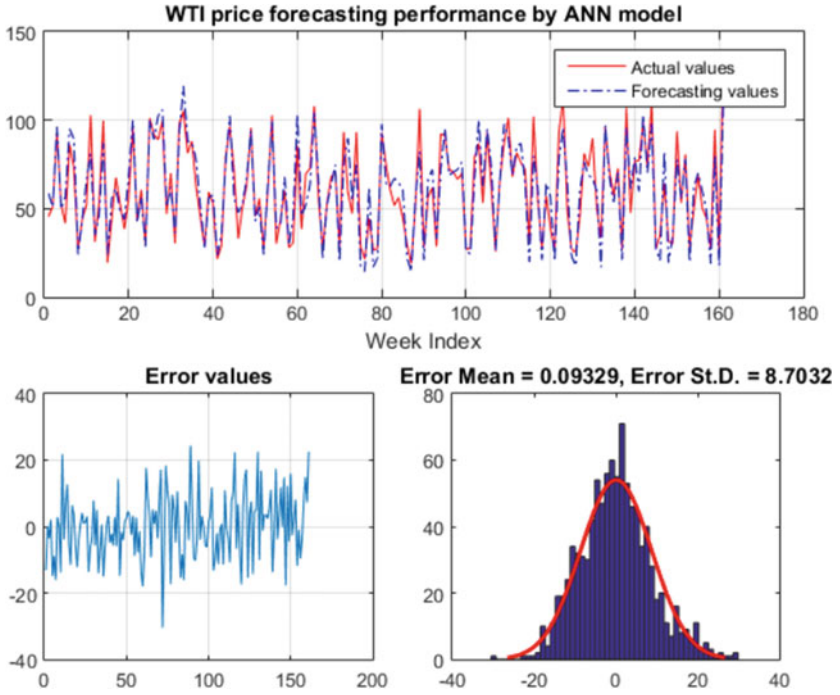


Fig. 8 WTI price forecasting performance by ANN model

form, $f(n)=2/(1+e^{-2n})-1$, is utilized; while a linear function is used from the hidden layer to the output layer. The parameters for back-propagation were set as follows: the learning and momentum rates were 0.4 and 0.3, respectively. For proposed ANFIS-BBO forecasting model, the parameters were set as follows: population size $p_s = 100$; maximum immigration rate $I = 1$; the maximum emigration rate $E = 1$; mutation probability: $m_{\max} = 0.005$; $\varepsilon = 0.1$; *limit* $= 100$; and $G = 500$.

The models were implemented in the MATLAB environment (Matlab 2015a). The simulation results were then obtained and presented in Table 1 and Figs. 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16. Due to the paper limit, Figs. 7, 8, 9, 10, 11, 12, 13 and 14 only present the values from August 2015 to September 2018. The time series of actual and forecasting values of WTI and Brent price obtained using four models including ANFIS-BBO based model, ANFIS-based model, ANN-based model and linear regression method are compared in Figs. 15 and 16. The nearly perfect agreement between the trends in the plots of the time series of actual and forecasting values in the figure, suggest that the ANFIS-BBO based model is the most suitable model.

The performance criteria R, MAE, RMSE, RAE and RRSE obtained by ANFIS-BBO based model for Brent price were calculated as 0.9865, 3.4741, 5.0983, 13.2955% and 0.9768 16.4517%, WTI price were calculated as 0.9828, 3.4626,

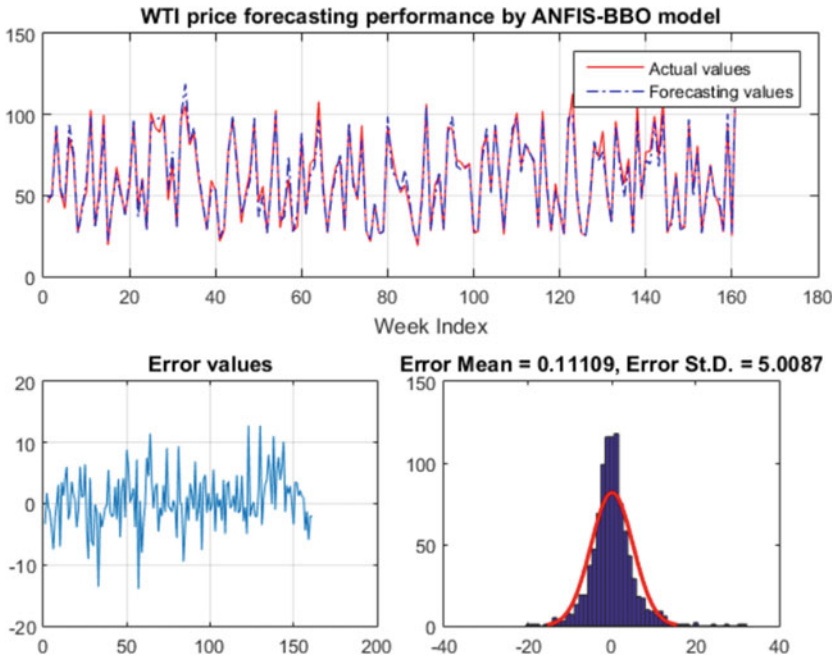


Fig. 9 WTI price forecasting performance by ANFIS-BBO model

5.0095, 15.0601% and 18.5077% and for R, MAE, RMSE, RAE and RRSE, respectively. Theoretically, a forecasting model is accepted as ideal when MAE, RMSE, RAE and RRSE are small, and R is close to 1. It is very clear from Table 1 that the ANFIS-BBO-based model has smaller errors as well as a bigger R than those of the ANN-based model, ANFIS-based model and linear regression method. The performance criteria indicate that the assessed result is highly correlated and precise.

The comparisons between actual values and forecasting values obtained by developed models are also shown in Figs. 15 and 16. The figure presents the scatter diagrams that illustrate the degree of correlation between forecasting values and actual values. An identity line was drawn as a reference. In this figure, the identity line represents that the two sets of data are identical. The more the two datasets agree, the more the points tend to concentrate in the vicinity of the identity line. It may be observed that the most forecasting values are very close to the actual values. This indicates a sound agreement between ANFIS-BBO based model forecasting and the actual values.

Based on the obtained results, it can be inferred that the proposed ANFIS-BBO based model can be used to forecast crude oil price. The ANFIS-BBO based model outperformed the ANN-based model, ANFIS based model and linear regression method, and the results showed that its forecasting outcome is more accurate and reliable. Hence, the ANFIS-BBO based model is acceptable and good enough to serve as a predictor of oil price.

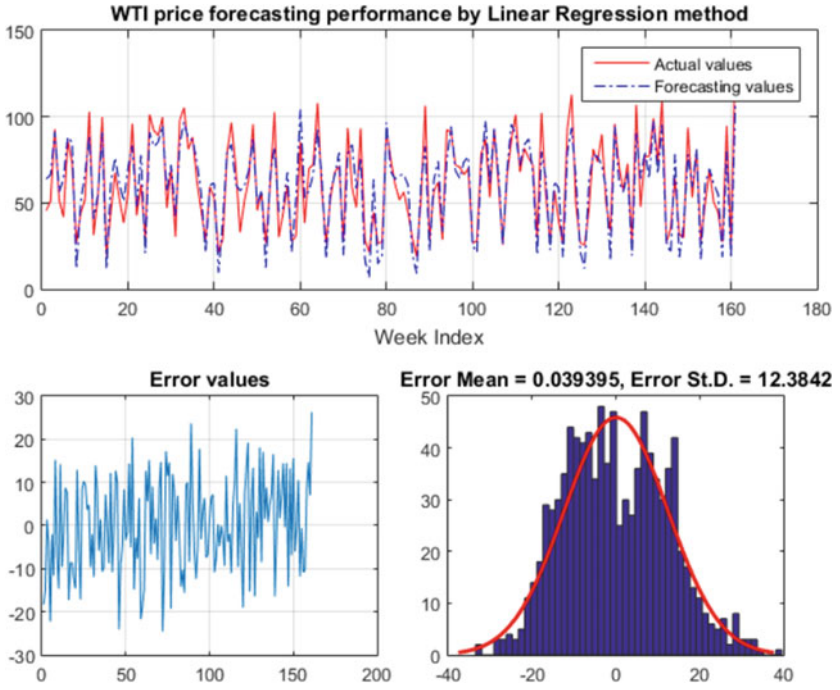


Fig. 10 WTI price forecasting performance by Linear Regression method

6 Conclusions

This study has developed different oil price forecasting models to forecast crude oil price in the international market including WTI and Brent. Among these developed model, a novel model based on the combination of MI theory, fuzzy logic, neural network, and BBO algorithm has been proposed. The results clearly demonstrated the superior forecasting performance of ANFIS-BBO model. It is concluded that MI-ANFIS-BBO can be utilized for crude oil price prediction. These findings also demonstrated the remarkable advantage of the BBO in training neural network. The results of the present study also show that using heuristic algorithms in optimizing parameters of a neural network can enhance the performance of a neural network-based model. This work can thus make a contribution to the development of selecting the best forecasting model for other purposes. For the future research recommendation, this study will be better if the proposed model will be integrated with more economic, social, and political indicators as inputs of the model to improve the quality of forecasting.

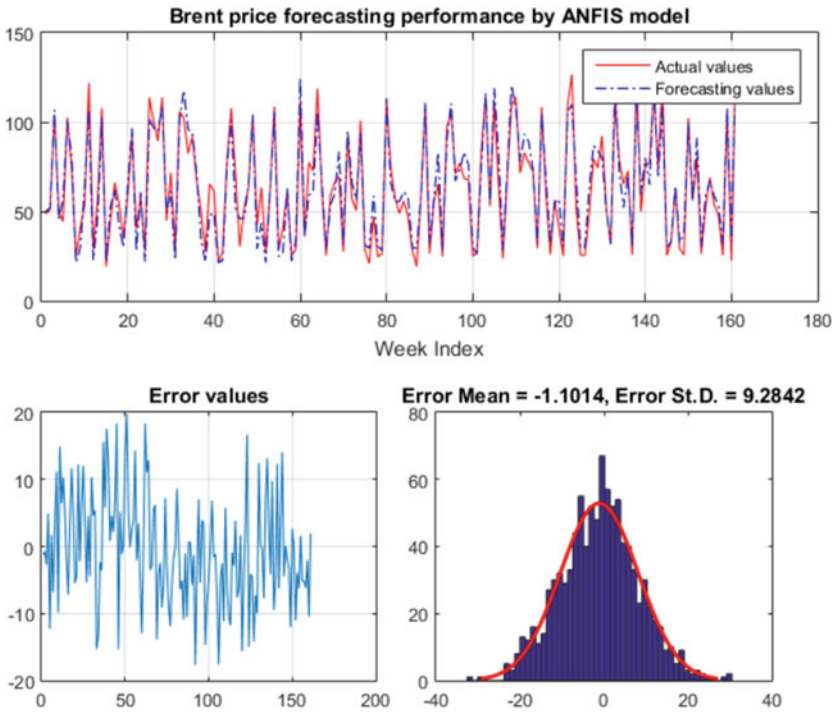


Fig. 11 Brent price forecasting performance by ANFIS model

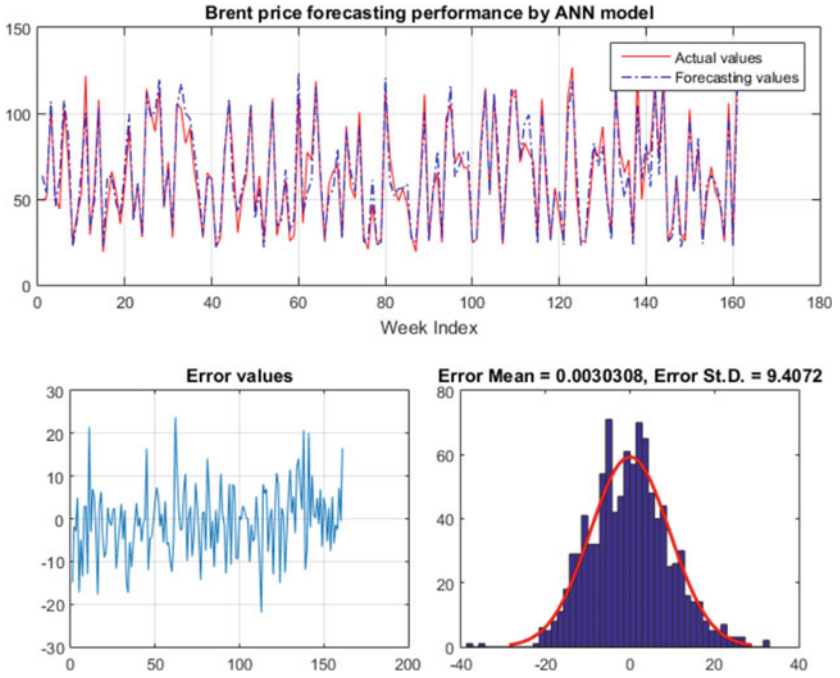


Fig. 12 Brent price forecasting performance by ANN model

Fig. 13 Brent price forecasting performance by ANFIS-BBO model

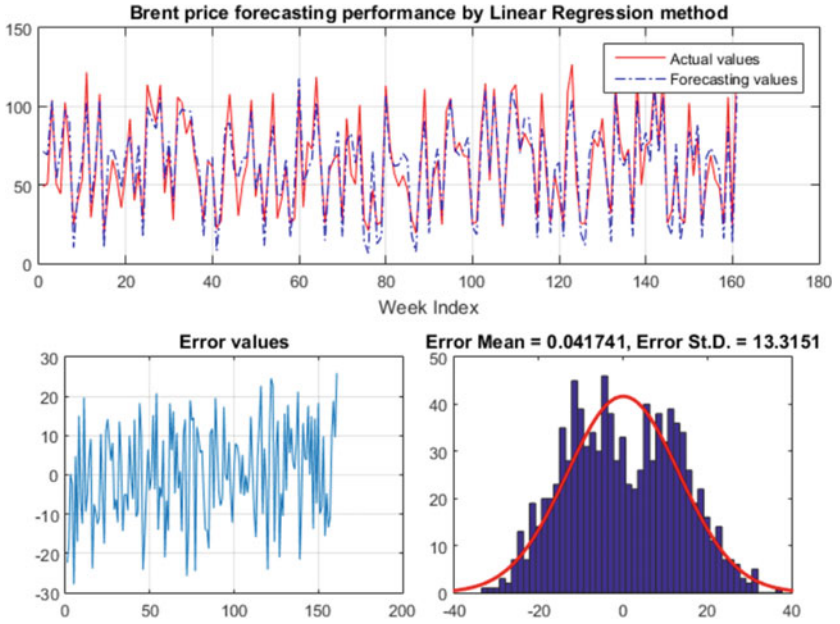


Fig. 14 Brent price forecasting performance by Linear Regression method

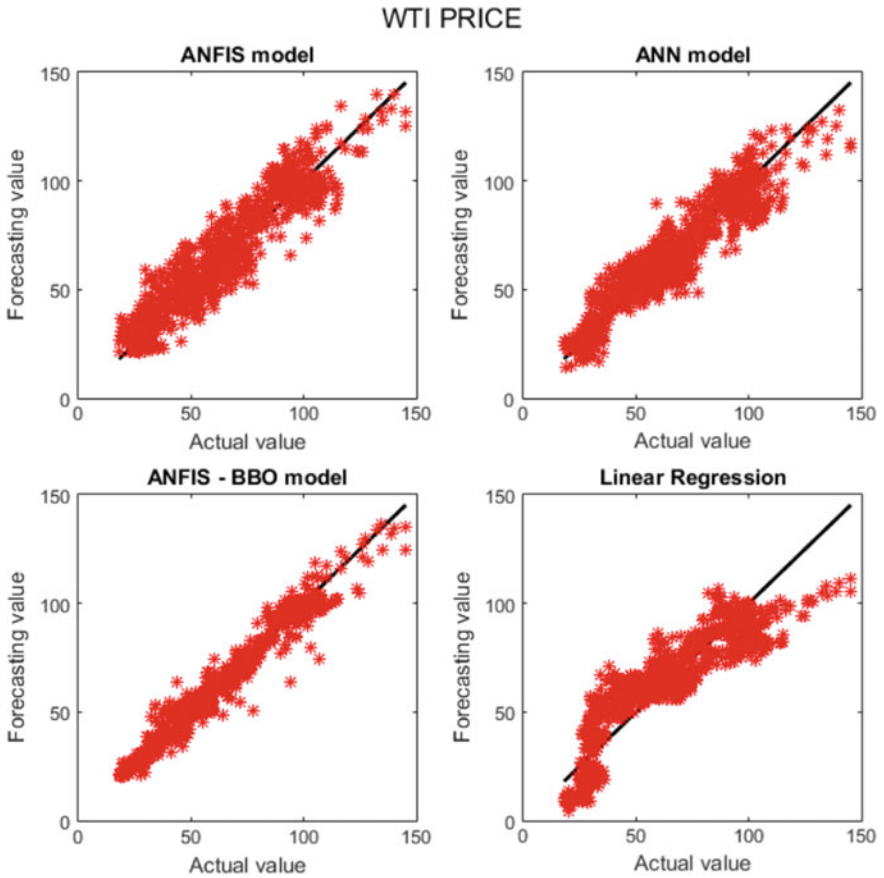


Fig. 15 Comparison between actual and forecasting price values of different models for WTI

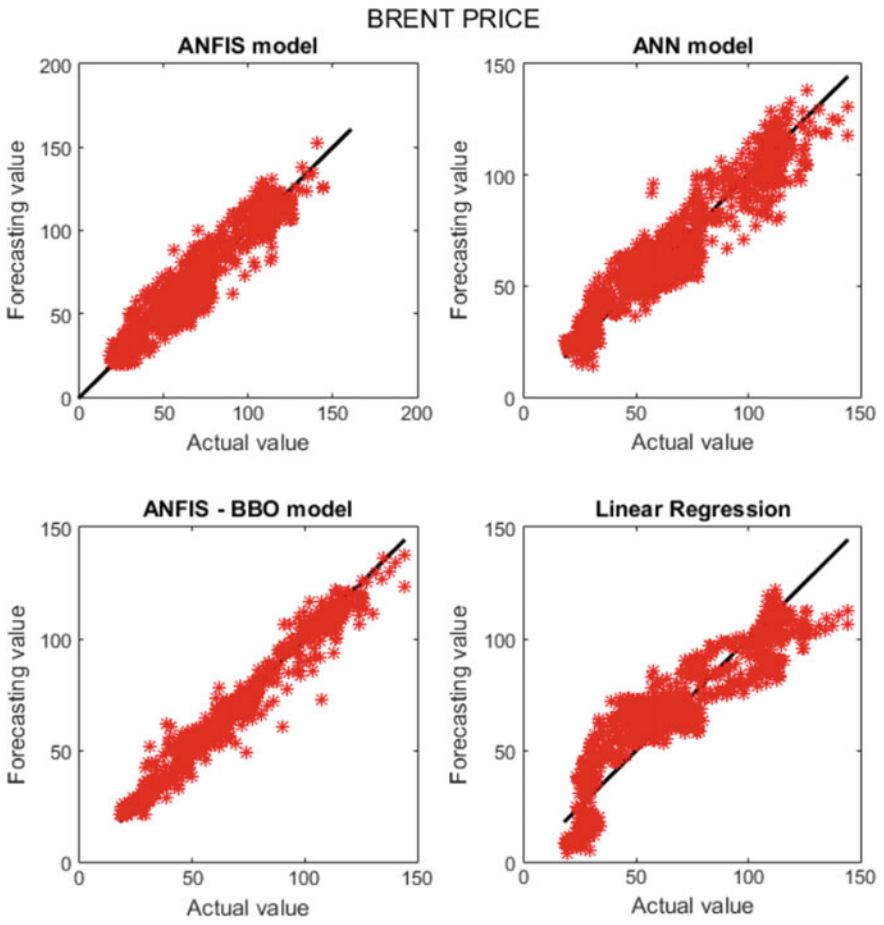


Fig. 16 Comparison between actual and forecasting price values of different models for Brent

References

1. Ata R, Kocyigit Y (2010) An adaptive neuro-fuzzy inference system approach for prediction of tip speed ratio in wind turbines. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2010.02.068
2. Azadeh A, Saberi M, Anvari M, Azaron A, Mohammadi M (2011) An adaptive network based fuzzy inference system-genetic algorithm clustering ensemble algorithm for performance assessment and improvement of conventional power plants. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2010.08.010
3. Bashiri Behmiri N, Pires Manso JR (2013) Crude oil price forecasting techniques: a comprehensive review of literature. SSRN Electron J. https://doi.org/10.2139/ssrn.2275428
4. Baumeister C (2014) The art and science of forecasting the real price of oil. Bank Can Rev 2014(Spring):21–31
5. Beckers B, Beidas-Strom S (2015) Forecasting the nominal brent oil price with VARs-one model fits all? In: IMF working papers. https://doi.org/10.5089/9781513524276.001

6. Bopp AE, Lady GM (1991) A comparison of petroleum futures versus spot prices as predictors of prices in the future. Energy Econ. https://doi.org/10.1016/0140-9883(91)90007-M
7. Buragohain M, Mahanta C (2008) A novel approach for ANFIS modelling based on full factorial design. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2007.03.010
8. Chen H, Liao H, Tang BJ, Wei YM (2016) Impacts of OPEC's political risk on the international crude oil prices: an empirical analysis based on the SVAR models. Energy Econ. https://doi.org/10.1016/j.eneco.2016.04.018
9. Doğan E, Akgüngör AP (2013) Forecasting highway casualties under the effect of railway development policy in Turkey using artificial neural networks. Neural Comput Appl. https://doi.org/10.1007/s00521-011-0778-0
10. Erzin Y, Gul T (2014) The use of neural networks for the prediction of the settlement of one-way footings on cohesionless soils based on standard penetration test. In: Neural computing and applications, vol 24. https://doi.org/10.1007/s00521-012-1302-x
11. Gao S, Lei Y (2017) A new approach for crude oil price prediction based on stream learning. Geosci Front. https://doi.org/10.1016/j.gsf.2016.08.002
12. Haddad OB, Hosseini-Moghari S-M, Loáiciga HA (2015) Biogeography-based optimization algorithm for optimal operation of reservoir systems. J Water Resour Plan Manage 142(1):4015034
13. Hamdi M, Aloui C (2015) Forecasting crude oil price using artificial neural networks: a literature survey. Econ Bull
14. Huang N, Hu Z, Cai G, Yang D (2016) Short term electrical load forecasting using mutual information based feature selection with generalized minimum-redundancy and maximum-relevance criteria. Entropy. https://doi.org/10.3390/e18090330
15. Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice Hall
16. Jang JSR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst, Man Cybern. https://doi.org/10.1109/21.256541
17. Karaboga D, Kaya E (2018) Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey. Artif Intell Rev. https://doi.org/10.1007/s10462-017-9610-2
18. Messai N, Thomas P, Lefebvre D, El Moudni A (2002) A neural network approach for freeway traffic flow prediction. In: Proceedings of the international conference on control applications, vol 2, pp 984–989. IEEE
19. Metin Ertunc H, Hosoz M (2008) Comparative analysis of an evaporative condenser using artificial neural network and adaptive neuro-fuzzy inference system. Int J Refrig. https://doi.org/10.1016/j.ijrefrig.2008.03.007
20. Mirjalili S (2019) Biogeography-based optimisation. Stud Comput Intell. https://doi.org/10.1007/978-3-319-93025-1_5
21. Mirmirani S, Cheng Li H (2004) A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil. In: Applications of artificial intelligence in finance and economics, pp 203–223. Emerald Group Publishing Limited
22. Morana C (2001) A semiparametric approach to short-term oil price forecasting. Energy Econ. https://doi.org/10.1016/S0140-9883(00)00075-X
23. Moshiri S, Foroutan F (2006) Forecasting nonlinear crude oil futures prices. Energy J. https://doi.org/10.5547/ISSN0195-6574-EJ-Vol27-No4-4
24. Nauck D, Klawonn F, Kruse R (1997) Foundations of neuro-fuzzy systems. Wiley
25. Pal SK, Mitra S (1999) Neuro-fuzzy pattern recognition: methods in soft computing. Wiley
26. Santana ÁL, Conde GB, Rego LP, Rocha CA, Cardoso DL, Costa JC, Bezerra UH, Francês CRL (2012) PREDICT—Decision support system for load forecasting and inference: a new undertaking for Brazilian power suppliers. Int J Electr Power Energy Syst. https://doi.org/10.1016/j.ijepes.2011.12.018
27. Shabri A, Samsudin R (2014) Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model. Math Prob Eng. https://doi.org/10.1155/2014/201402

28. Simon D (2008) Biogeography-based optimization. IEEE Trans Evol Comput. https://doi.org/10.1109/TEVC.2008.919004
29. Simon D (2011) A probabilistic analysis of a simplified biogeography-based optimization algorithm. Evol Comput. https://doi.org/10.1162/EVCO_a_00018
30. Singh R, Kainthola A, Singh TN (2012) Estimation of elastic constant of rocks using an ANFIS approach. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2011.09.010
31. Sugeno M (1985) An introductory survey of fuzzy control. Inf Sci. https://doi.org/10.1016/0020-0255(85)90026-X
32. Takagi H, Hayashi I (1991) NN-driven fuzzy reasoning. Int J Approx Reason. https://doi.org/10.1016/0888-613X(91)90008-A
33. Takagi T, Sugeno M (1984) Derivation of fuzzy control rules from human operator's control actions. IFAC Proc Ser
34. Tamjidy M, Paslar S, Baharudin BTHT, Hong TS, Ariffin MKA (2015) Biogeography based optimization (BBO) algorithm to minimise non-productive time during hole-making process. Int J Prod Res. https://doi.org/10.1080/00207543.2014.965356
35. Wei M, Bai B, Sung AH, Liu Q, Wang J, Cather ME (2007) Predicting injection profiles using ANFIS. Inf Sci. https://doi.org/10.1016/j.ins.2007.03.021
36. Yu L, Wang S, Lai KK (2008) Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Econ. https://doi.org/10.1016/j.eneco.2008.05.003
37. Yuan J, Mills K (2005) A cross-correlation-based method for spatial-temporal traffic analysis. Perform Eval. https://doi.org/10.1016/j.peva.2004.11.003
38. Yusof N, Bahiah N, Shahizan M, Chun Y (2012) A concise fuzzy rule base to reason student performance based on rough-fuzzy approach. Fuzzy Infer Syst Theory Appl. https://doi.org/10.5772/37773

Quang Hung Do (in Vietnamese: Đ˜ô Quang Hưng), Ph.D., is a faculty member of University of Transport Technology (Vietnam). He was a researcher and a lecturer at Feng Chia University (Taiwan). He has published more than 50 papers in international journals and conferences. He is a reviewer for several SCI-indexed journals, including Applied Soft Computing, Journal of the Operational Research Society. He also served as a program committee member for several international conferences such as ACITY-2017, CSITY-2018, CSEN-2018 and ITCA-2019. His primary field of *research interest* is *Artificial Intelligence* (*AI*) and its *applications in business, management, and engineering*

Obtaining More Specific Topics and Detecting Weak Signals by Topic Word Selection

Laura Kölbl and Michael Grottke

Abstract With topic modeling methods, such as Latent Dirichlet Allocation (LDA), we can find topics in large text collections. To efficiently employ this information, there is a need for a method that automatically analyzes the topics with respect to their usefulness for applications like the detection of new innovations. This paper presents a novel method to automatically evaluate topics produced by LDA. The new approach puts the focus on finding topics with topic words that are not only coherent, but also specific. By using the documents associated with each word to calculate background topics, a baseline can be set for each topic word that helps assess whether its context fits the topic well. Experiments indicate that the resulting topics are more manageable in terms of their interpretability. Moreover, we show that the approach can be used to detect weak signals.

Keywords Text mining · Topic modeling · Weak signals · Topic coherence

1 Introduction

A major drawback in topic modeling is the (lack of) interpretability of topics. Especially for a large number of topics, human evaluation of models and topics is a time-consuming task. Since the results of such a human evaluation do not correlate well with automatic measures like perplexity and held-out likelihood [1], there is a need for more adequate metrics. State-of-the-art topic coherence measures perform fairly well [2], but they do not consider how specific and meaningful a word is.

Depending on the task it can be crucial that the topics are very specific. Chang et al. [1] found that topic coherence (as measured by humans) declines with an increasing

L. Kölbl (✉) · M. Grottke
Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany
e-mail: laura.koelbl@fau.de

M. Grottke
e-mail: michael.grottke@fau.de; michael.grottke@gfk.com

M. Grottke
GfK SE, Global Data Science, Nürnberg, Germany

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_12

number of topics. However, fine-grained topics are important for application tasks such as emerging trend detection. Therefore, there is a need for an approach capable of identifying good topics among the many topics found by LDA.

In this paper, we assess the quality of a topic in terms of both coherence and specificity by using context information from the document corpus.

2 Related Work

2.1 Latent Dirichlet Allocation

LDA [3] is a generative probabilistic model of a document collection. It assumes that each document is a mixture of latent topics, which in turn are mixtures of words. The input is a corpus $\mathcal{C}$ consisting of D documents. $\mathcal{V}$ is the vocabulary containing all the different words present in this corpus. Document $d \in \{1, \ldots, D\}$ consists of N_d words.

The topic-word distributions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ of the K topics are drawn from a Dirichlet distribution with hyperparameter vector $\boldsymbol{\beta}$, while the topic distributions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D$ of the D documents are sampled from a Dirichlet distribution with hyperparameter vector $\boldsymbol{\alpha}$. The documents are thus assumed to be generated using Algorithm 1.

Algorithm 1: Document generation assumed by LDA

$\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}), k \in 1, \ldots, K$;
for *the* d^{th} *document,* $d \in \{1, \ldots, D\}$, **do**
  $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$;
  for *the* n^{th} *word in document* d, $n \in \{1, \ldots, N_d\}$, **do**
    Choose a topic $z_{d,n} \sim$ Multinomial($\boldsymbol{\theta}_d$);
    Choose a word $t_{d,n} \sim$ Multinomial($\boldsymbol{\phi}_{z_{d,n}}$);
  end
end

2.2 Coherence Measures

There are different approaches to the handling of topics modeled with LDA that are related to our work.

The paper by Lau et al. [4] about selecting topic words as well as subsequent papers on this subject matter [5] have aimed at finding the best label for a topic.

Recent work by Alokaili [6] is concerned with topic word reranking; the aim of this paper is to enhance the quality of the topics by reordering the topic words.

Although topic reranking seems closely related to our work, the methods used in the paper do not incorporate any semantic context information, but they rely for example on term frequencies or probabilities. We will use word selection in order to find specific topics with an approach that emphasizes context information; therefore, we will employ coherence measures as a benchmark in our work. Topic coherence captures the internal relatedness of the topic.

Research on topic coherence measures has mainly focused on topic words. In the literature, usually the top-W words (where $W = 5$ or $W = 10$), evaluated by their topic-word probabilities, are defined to form the set of topic words for topic k, $\mathcal{W}_k = \{w_{k,1}, \ldots, w_{k,W}\}$. Word topic coherence can be determined by comparing pairs of these words [7], [8], or by comparing words with word subsets, as done by Rosner et al. [9] and Röder et al. [2], who considered one-all, one-any and one-many comparisons.

According to Röder et al. [2], the measure that showed the best results for topic coherence compared with human evaluation was the C_V measure, which is based on comparing each topic word with the entire topic word set. Calculating how $w_{k,i}$ is supported by $\mathcal{W}_k$ results in a vector $\boldsymbol{v}_{k,i}$ whose j^{th} element represents the comparison of $w_{k,i}$ with $w_{k,j}$ using the normalized pointwise mutual information (NMPI):

$$v_{k,i,j} = \text{NMPI}(w_{k,i}, w_{k,j}) = \frac{\ln \frac{P(w_{k,i}, w_{k,j}) + \epsilon}{P(w_{k,i}) \cdot P(w_{k,j})}}{-\ln(P(w_{k,i}, w_{k,j}) + \epsilon)}.$$

The small constant ϵ is added in the logarithms to avoid problems with zero probabilities. These (joint) word probabilities are computed using a Boolean sliding window.

While $\boldsymbol{v}_{k,i}$ could be used as a direct confirmation measure for topic word $w_{k,i}$, the C_V metric is based on comparing this vector with vector $\boldsymbol{v}_k^*$, representing the support of word set $\mathcal{W}_k$ with each word in $\mathcal{W}_k$. The jth element of this vector can easily be calculated as

$$v_{k,j}^* = \sum_{i=1}^{W} v_{k,i,j}.$$

An indirect confirmation measure for topic word $w_{k,i}$ is then calculated as the cosine vector similarity between $\boldsymbol{v}_{k,i}$ and $\boldsymbol{v}_k^*$,

$$\psi_{k,i} = Sim_{cos}(\boldsymbol{v}_{k,i}, \boldsymbol{v}_k^*),$$

and the C_V metric for topic k is the arithmetic average of the W measures $\psi_{k,1}, \ldots, \psi_{k,W}$.

Our aim in this paper is to add information about the semantic context of the words to the coherence measure.

Recent work by Korencic et al. [10] has suggested to use document-based coherence measures for detecting concrete topics. The authors computed document-based coherence by selecting the top documents for a topic, and they evaluated their coher-

ence by calculating for example their cosine distance. The paper highlighted that some topics showed low word coherence although the documents were very similar. Document coherence approaches topic coherence from a very different and interesting angle.

In the following, we present a novel approach, which detects concrete topics by finding specific topic words.

3 Proposed Approach

AlSumait et al. [11] introduced the idea of comparing each topic with "junk topics", which are word or document distributions that are inferred from the whole corpus. They used the similarity of a topic to these junk topics as a measure of topic quality.

Based on this idea, we also make use of background topics, but unlike in the approach by AlSumait et al. we define background topics that are word-specific in order to detect topic words that are tightly connected with a particular topic.

First, we run an LDA with K topics. We consider the top-W words of topic k to form its set of topic words $\mathcal{W}_k$. For each of these topic words $w_{k,i}$, we build the background corpus $\mathcal{C}_{k,i}$, i.e., the corpus consisting of all the documents that contain word $w_{k,i}$. From background corpus $\mathcal{C}_{k,i}$ a background topic $b_{k,i}$ is calculated by running an LDA with the topic number parameter set to one.

Using the Jensen-Shannon divergence (JSD), we can compute a similarity measure comparing the word distribution of background topic $b_{k,i}$ with the word distribution of any of the K topics from the original LDA:

$$\text{BGM}\,(k', b_{k,i}) = 1 - \text{JSD}\,(\boldsymbol{\phi}_{k'}, \boldsymbol{\phi}_{b_{k,i}}), \quad k' = 1, \ldots, K.$$

If a word $w_{k,i}$ featuring in the topic list $\mathcal{W}_k$ is specific to this topic, then comparing $\boldsymbol{\phi}_k$ with $\boldsymbol{\phi}_{b_{k,i}}$ should result in a relatively high value. On the contrary, the background topic of any word that is not specific to a topic will be unlikely to achieve the highest similarity with this one topic. Rather, if the context of the word is not closely related to the topic, then it might fit some other topic well.

Whether a topic word is to be considered specific also depends on the corpus studied. For example, a topic word like "health" could be deemed specific in a dataset about robotics but unspecific in a dataset about healthcare. We might thus be inclined to keep topic word $w_{k,i}$ in $\mathcal{W}_k$ only if $\text{BGM}(k, b_{k,i})$ is the largest similarity value obtained.

However, even a topic-specific word might be relevant for more than one topic. In our algorithm, we therefore use the background indicator (BGI) parameter to decide whether or not a word is to be dropped from the topic list. Arranging the similarities $\text{BGM}(k', b_{k,i})$, $k' \in \{1, \ldots, K\}$, in descending order, the BGI defines the rank position up to which the word is still accepted. The most restrictive choice $\text{BGI} = 1$ would lead us to the approach laid out in the previous paragraph, demanding

that the background topic $b_{k,i}$ attains the highest similarity when comparing it to topic k.

Algorithm 2: BGI Algorithm

Input: $\mathcal{C}$, K, BGI
Run LDA with K topics on document corpus $\mathcal{C}$;
for *the* k^{th} *topic,* $k \in \{1, \ldots, K\}$, **do**
  for *the* i^{th} *topic word in* $\mathcal{W}_k$, $i \in \{1, \ldots, W\}$, **do**
    Create background corpus $\mathcal{C}_{k,i}$;
    Obtain background topic $b_{k,i}$ by running LDA with 1 topic on $\mathcal{C}_{k,i}$;
    Arrange similarities BGM$(1, b_{k,i}), \ldots,$ BGM$(K, b_{k,i})$ in descending order;
    if *rank of* $BGM(k, b_{k,i}) > BGI$ **then**
      drop $w_{k,i}$ from $\mathcal{W}_k$
    end
  end
end

For a larger BGI value, such as 2, the algorithm drops fewer words from the list of topic words, because it also keeps those for which topic k only attains a lower (e.g., the second-largest) similarity value. After all, the ultimate aim of our approach is not to separate the contents of the various topics, but to assess the topic quality and to enhance the interpretability of the topics. The whole approach is depicted in Algorithm 2.

By selecting the topic words, our approach is implicitly selecting topics as well: if all of the topic words of a topic are dropped, then the topic can be considered unspecific.

Although this approach seems computationally intensive, there is no need to calculate all the similarities BGM$(1, b_{k,i}), \ldots,$ BGM$(K, b_{k,i})$ for each topic word: as soon as BGI topics have a better rating than topic k, the topic word can be dropped.

4 Topic Word Selection

4.1 Experimental Set-Up

In the following, we describe an experimental set-up to show the applicability of the approach as a topic word selection approach.

We are using two data sets to evaluate the performance of our approach: The first data set is a technology dataset publicly available at webhose.io. It comprises 22,292 news and blog texts from September 2015 on technology scraped from the web in

that month.[1] The second data set consists of 2669 abstracts of scientific articles on autonomous vehicles[2] from the years 2017 and 2018.

We execute LDA[3] with the entries of the hyperparameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ set to 0.5 and 0.01, respectively, the values used by Rosner et al. [9]. The topic parameter K is set to 50. We calculate the results for the BGI parameter set to 1 and 2. $W = 10$ topic words are considered for the BGI approach and the C_V measure; for the latter, we use $\epsilon = 10^{-12}$ as suggested by Röder et al. [2].

For BGI $= 1$, our approach selects 15 topics for the technology data set and 18 topics for the autonomous vehicles data set. In the following, we will present the top-10 results for each data set. The rank rk_B of each topic is determined by the number of words selected by the BGI algorithm; in case of ties we are using mid-ranks.

The C_V measure is the one most similar to our approach, because it also uses information from the documents associated with each topic word, and it accounts for indirect confirmation between the topic words. In our results tables, we therefore provide the values and ranks based on the C_V measure for comparison.

4.2 Topic Word Selection Results

Table 1 shows the results for the webhose technology data set. Words in bold (in italics) were chosen with the BGI parameter set to 1 (2).

As intended, our approach tends to pick words specific for one topic (such as "gene" and "dna" in the seventh topic) over more general terms ("paper", "science"). Terms that are only slightly connected to a topic are rejected (e.g., "year" and "expect" for the eighth topic). This is not the case for all the 10 highest-ranking topics according to C_V. At the bottom of the table, we show the topics ranked 6–8 by C_V. While these topics have a relatively high word coherence, they comprise many general terms.

From the results it also becomes clear that named entities are often selected by the approach. For example, "Kumari" and "Biswas" are the last names of scientists who authored papers on secure authentification, which are cited in the news articles. Other examples include "medigus", "nasa", and "apple". The named entities in the ninth topic point to the 2012 shooting of the imam Obid-kori Nazarov in Sweden. Obviously, this topic has been formed based on non-technology-related articles contained in the document corpus, and the BGI algorithm has correctly identified the topic words as topic-specific.

Table 2 shows the outcomes for the autonomous vehicles data set.

Again, these results confirm that the words are selected based on how specific they are for each topic, in relation to the other topics. The term "time" is rather

[1] Publicly available at https://www.webhose.io/datasets; retrieved on August 27th, 2018.

[2] All the paper data sets were downloaded from Scopus API between January 4th, 2019 and February 4th, 2019 via http://www.api.elsevier.com and http://www.scopus.com.

[3] In all our experiments we used common preprocessing techniques such as stemming and custom stopword lists to enhance the topic quality.

Table 1 Topic selection results for the technology data set; the first column rk_B shows the ranking according to the BGI approach and the last column rk_{C_V} shows the ranking according to the C_V measure; topic words in bold (in italics) were selected with the BGI parameter set to 1 (2)

rk_B	Topic Words, **BGI = 1**, *BGI = 2*, BGI > 2	C_V	rk_{C_V}
1	**'cloud', 'manag', 'custom', 'solut', 'softwar', 'platform', 'servic', 'data', 'busi', 'technolog'**	0.647	10
2	**'nasa', 'earth', 'ocean', 'sea', 'temperatur', 'satellit', 'water'**, *'imag'*, 'space', 'smartwatch'	0.602	13
3	**'oneplu', 'redmi', 'processor', 'batteri', 'display', 'camera', 'samsung'**, 'smartphon', 'note', 'phone'	0.741	4
4.5	**'kumari', 'biswa', 'cryptanalysi', 'telecar', 'syst', 'crossref'**, *'authent'*, 'agreement', 'medicin', 'doi'	0.926	1
4.5	**'uzbek', 'teacher', 'indian', 'friday', 'student', 'school'**, *'distribut'*, 'technolog', 'industry', 'presid'	0.660	9
6	**'physician', 'clinic', 'patient', 'surgic', 'medic'**, *'health', 'care', 'healthcar'*, 'procedur', 'announc'	0.813	2
7	**'gene', 'dna', 'scientist', 'cell', 'brain'**, *'professor', 'paper'*, 'univers', 'science', 'human'	0.754	3
8	**'siri', 'iphon', 'appl', 'pro'**, *'box', 'watch', 'app', 'ipad'*, 'year', 'expect'	0.695	5
9	**'kupaisinov', 'tashkent', 'zhukovski', 'nazarov'**, 'amin', 'suspect', 'attack', 'imam', 'told', 'polic'	0.504	33
10	**'forwardlook', 'medigu'**, *'statement', 'futur', 'risk'*, 'includ', 'code', 'secur', 'releas', 'compani'	0.504	32
–	'famili', 'children', 'women', 'parent', 'age', 'child', 'tree', 'kid', 'young', 'men'	.690	6
–	*'investor', 'bank', 'capit'*, 'compani', 'invest', 'busi', 'billion', 'fund', 'firm', 'financi'	0.662	7
–	'million', 'year', 'quarter', 'revenu', 'loss', 'end', 'total', 'compar', 'net', 'share'	0.661	8

unspecific for the sixth topic about driver assistance systems or the ninth topic about lane changing; however, it is much more closely related to the third topic dealing with public transport (e.g., the departure times and the travel times experienced).

"Thing", a highly unspecific term by definition, was nevertheless chosen for the second topic, because in this data set the semantic context of the word is deeply connected with the collective term "internet of things".

Topic ten is concerned with traffic flow simulations. Microscopic traffic flow simulations are models used for simulating the traffic dynamics of a single vehicle.

Table 2 Topic selection results for the autonomous vehicles data set; the first column rk_B shows the ranking according to the BGI approach and the last column rk_{C_V} shows the ranking according to the C_V measure; topic words in bold (in italics) were selected with the BGI parameter set to 1 (2)

rk_B	Topic words, **BGI = 1**, *BGI = 2*, BGI > 2	C_V	rk_{C_V}
2	**'camera', 'detect', 'map', 'imag', 'vision', 'video', 'match', 'visual', 'point', 'target'**	0.666	6
2	**'busi', 'innov', 'world', 'internet', 'iot', 'digit', 'technolog', 'thing', 'emerg', 'creat'**	0.725	3
2	**'travel', 'citi', 'public', 'transport', 'servic', 'urban', 'user', 'choic', 'demand', 'time'**	0.695	4
4	**'nonlinear', 'program', 'constraint', 'solv', 'comput', 'dynam', 'set', 'uncertainti', 'solut'**, 'time'	0.620	7
5	**'energi', 'emiss', 'electr', 'fuel', 'batteri', 'pollut', 'power', 'hybrid'**, *'air', 'reduct'*	0.818	2
6	**'drive', 'autom', 'driver', 'task', 'situat', 'assist', 'steer', 'condit'**, *'level'*, 'time'	0.689	5
7.5	**'neural', 'deep', 'recognit', 'learn', 'machin', 'train', 'artifici'**, *'intellig', 'task'*, 'network'	0.830	1
7.5	**'navig', 'filter', 'posit', 'error', 'path', 'updat', 'plan'** *'dynam', 'autonom'*, 'environ'	0.506	13
9.5	**'penetr', 'lane', 'connect', 'rate', 'vehicl'**, *'autom', 'mix'*, 'time', 'level', 'commun'	0.529	9
9.5	**'flow', 'microscop', 'traffic', 'fundament'**, *'highway', 'congest'*, 'desir', 'condit', 'evid', 'assign'	0.501	15
13	**'stabil', 'theoret', 'paramet'**, *'trajectori', 'scheme', 'platoon', 'refer', 'dynam', 'numer'*, 'robust'	0.553	8
14	**'fatal', 'crash', 'accid'**, *'prevent'*, 'road', 'safeti', 'option', 'motor', 'rise', 'number'	0.527	10

5 Weak Signal Detection

5.1 Terminology

In a second experiment, we show how to employ the proposed approach for finding weak signals. Weak signal detection is thereby located in the subject area of corporate foresight, where topic models have already been shown to be a useful tool to find new innovation opportunities [12].

A weak signal is defined as a new event with the capability of having an impact on future events [13]. This makes weak signals hard to pinpoint, because their status as a weak signal does not depend on whether or not they will *actually* have an impact later on, but rather on their *potential ability* to do so. In other words, a weak signal is an early indicator of change.

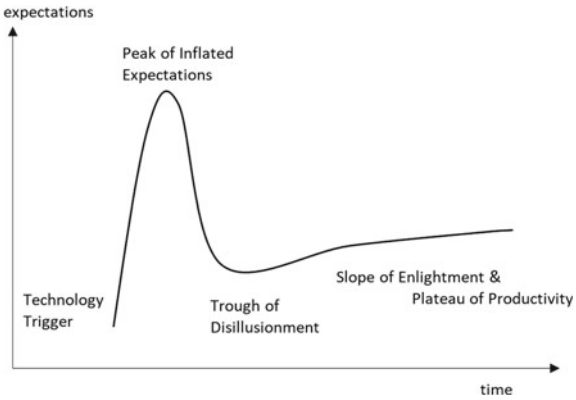


Fig. 1 Gartner hype cycle [16]

This differentiates weak signals from trends; Saritas and Smith [14] define an (emerging) trend as follows: "Possible new trends grow from innovations, projects, beliefs or actions that have the potential to grow and eventually go mainstream in the future". Therefore, a weak signal can be seen as the possible root cause for a potential upcoming trend. A summary of papers on weak signal and trend detection is available in the systematic literature review by Mühlroth and Grottke [15].

The Gartner hype cycle presents the maturity of technologies in a graph that displays the expectations over time (see Fig. 1), and divides the development of emerging technologies into five phases: The first stage is the "technology trigger", where a technological breakthrough or a similar event draws a lot of attention. In the next phase, called the "peak of inflated expectations", excessive expectations are created about the new technology, that it ultimately cannot meet. These high expectations therefore sooner or later result in a "trough of disillusionment", which finally leads to phases of solid productivity and reasonable expectations ("slope of enlightment", "plateau of productivity") [16].

A weak signal can be seen to occur at the earliest phase of the hype cycle, as being or announcing a technology trigger. Gartner hype cycles are published every year for technology trends in general but also for specific areas like 3D printing and artifical intelligence. Technology innovations at early stages, between the technology trigger event and the peak, are displayed as "on the rise". Therefore, we consider a technology which appears on a hype cycle as "on the rise" and which was not contained in the hype cycle of the previous year as a sign of a newly-occurring weak signal.

5.2 Experimental Set-Up

In the experiment, we intend to retrace the weak signals on the gartner hype cycles on 3D printing from 1014, 2015 and 2016 using the BGI approach. To evaluate the

detection of weak signals, a data set on 3D printing and additive manufacturing was gathered.[4] It consists of 15,403 documents from the years 2010 through 2018.

The idea is to find occurrences of new words in the data set; if a new word is also specific for a topic, then it might hint at a weak signal. Therefore we consider a word-topic combination of such kind as a potential weak signal.

For example, to detect small, new topics with our approach, the initial LDA can be run with the data from the current year while the word selection can be applied using the data from the previous three years. Hereby, we drop all those words and topics that the BGI approach does not pick for BGI $= 2$. Then $w_{k,min}$ is defined as the topic word associated with the smallest number of documents from the respective year.

In our experiment, we applied the algorithm using the documents available at the end of the years 2014, 2015, and 2016. For each year, the algorithm was run for $K = 100$ and $K = 150$. Similar to comparable literature, such as Chang et al. [1], we thus increased K in steps of 50; unlike Chang et al. we started with $K = 100$ (rather than $K = 50$), to catch smaller topics. Due to the stochastic nature of the LDA, we calculated each model twice; i.e., the algorithm was run four times for each year. The LDA parameter settings were the same as in Sect. 4.

The topics were sorted in descending order by the number of documents connected with the respective word $w_{k,min}$. The top-10 topics of each algorithm run were considered as potential weak signals, and we compared them to topics that are newly marked as "on the rise" in the Gartner hype cycles 2015 and 2016 on 3D printing compared with the first Gartner hype cycle on 3D printing from 2014.[5]

To consider a potential weak signal to match a topic from the hype cycle, it was required that both the topic words and the documents associated with the selected words are connected with the hype cycle topic. The top-5 topic words were chosen according to the topic-word probabilities from the model, and the documents were chosen by using the selected word as a query on the documents of that respective year; to discard documents that are only loosely connected with the words (e.g., as part of an enumeration or example), we only count documents that have at least two occurences of the respective word as truly associated with the word and the topic.

Because of the topic word selection, small topics mainly consist of only a few but specific words. Thus sorting out the topics manually can be done much faster than with regular topics.

[4] All the paper data sets were downloaded from Scopus API between January 4th, 2019 and February 4th, 2019 via http://www.api.elsevier.com and http://www.scopus.com.

[5] www.gartner.com/doc/2803426/hype-cycle-d-printing; www.gartner.com/doc/3100228/hype-cycle-d-printing;
www.gartner.com/doc/3383717/hype-cycle-d-printing; viewed on February 15th, 2019.

5.3 Weak Signal Detection Results

The first occurrences of the hype cycle topics that were found are shown in Table 3. We found traces of 9 of the 12 topics that newly appeared in the hype cycles of 2015 and 2016.

The approach detected the bioprinting of organs in 2014 and 3D-printing-aided hip and knee implants in 2015, and first traces on the topic "3D printed drugs" could already be found in 2014. Also the hype cycle topic on Workflow Software of 2016 was yet traced in 2014, with the abbrevation "CAD" translating to "computer-aided design".

The topic about "wearables" was not detected per se, but it was possible to retrace early signs of research on the connection of 3D printing and biosensors in 2014.

The three Gartner hype cycle topics that have not been detected are nanoscale 3D printing (2016), 4D printing (2016), and sheet lamination (2016).

To show that the approach can detect the topics early in the data, the history of the words connected to the weak signals from Table 3 is provided in Table 4.

In 2015, when the topics "Consumable Products" and "Knee Implants" were detected by the approach, there were only 13 documents associated with the word "food" and merely 11 documents associated with the word "knee".

Table 3 Results for the weak signal detection approach; the table shows when the weak signal was first found, and the respective topic words selected by the BGI approach as well as the top-5 topic words

Gartner hype cycle topic	First found	Selected words	Top-5 topic words
Consumable Products, 2015	2015	'food'	'food', 'product', 'manufactur', 'technolog', 'chain'
Knee Implants, 2015	2015	'knee'	'knee', 'nativ', 'prostesi', 'asc' (stem cells), 'nasal'
Hip Implants, 2015	2015	'hip'	'cartilag', 'graft', 'scaffold', 'week', 'rat'
BP for Life Science R&D, 2015	2015	'market'	'market', 'life', 'science', 'emerg', 'sector'
Organ Transplants, 2015	2014	'organ'	'organ', 'engin', 'tissu', 'scaffold', 'biolog'
3D-Printed Drugs, 2016	2014	'tablet'	'drug', 'release', 'innov', 'form', 'deliveri'
Powder Bed Fusion, 2016	2015	'bed'	'melt', 'thermal', 'life', 'cycl', 'compon'
Workflow Software, 2016	2015	'software', 'cad'	'comput', 'cad', 'aid', 'subject', 'geometr'
Wearables, 2016	2014	'sensor'	'vascular', 'sensor', 'descript', 'uniqu', 'hardwar'

Table 4 Document history for the selected words; the table shows the absolute frequencies of documents that contain the respective word at least twice (bold type indicates when the topic was first found); the last row shows the total sum of documents in the corpus for the respective year

Selected words	2010	2011	2012	2013	2014	2015	2016	2017	2018
Food	0	1	1	2	0	**3**	16	24	38
Knee	0	6	0	4	2	**6**	7	21	12
Hip	0	2	3	0	0	**11**	14	22	31
Market	1	0	1	6	6	**15**	19	14	23
Organ	3	9	3	12	**18**	31	51	65	90
Tablet	1	0	0	0	**2**	7	5	14	22
Drug	4	7	12	13	17	**27**	37	71	136
Bed	1	2	1	2	5	**12**	17	43	60
Cad	3	4	6	9	20	**26**	31	20	48
Sensor	4	2	3	5	**10**	41	61	88	128
$\sum$ documents	208	299	350	535	913	1558	2474	3784	5282

We also added "drug" to the table as additional information, since "tablet" was the selected word for first occurrence of this weak signal in 2014 but it also appeared in 2015 as a weak signal with "drug" as a selected word. Furthermore, the word "drug" might be more representative for the topic "3D-Printed Drugs".

The table also shows the total sum of documents in the corpus for each year. This number noticeably increases over the years, which is explicable due to the increasing interest in 3D printing in general but also with the general increase in published papers per year. This should be considered before interpreting increases in publications on a certain topic.

6 Conclusion

In this paper, a new approach was proposed for finding specific topics by using context information from the topic words.

The results for the topic word selection demonstrate that the topics become more readable and on point. Another advantage of our approach over the benchmark is that it automatically selects the topics: there is no need to set a topics threshold. Although it is still necessary to set the number of topic words as a parameter, the BGI approach is less prone to influences of topic cardinality on the topic quality estimation [17] due to the word selection.

We have also shown that we can find new topics at very early stages of their appearance in the corpus. The results indicate that the approach might be able to detect weak signals for corporate foresight very early when applied to more extensive data sets.

Acknowledgements This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Program Concept Innovations for "Tomorrows Production, Services, and Work" (02K16C190) and managed by the Project Management Agency Forschungszentrum Karlsruhe (PTKA). The authors are responsible for the contents of this publication.

References

1. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Advances in neural information processing systems, vol 22. Vancouver, Canada, pp 288–296
2. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining. Shanghai, China, pp 399–408
3. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
4. Lau, JH, Newman D, Karimi S, Baldwin T (2010) Best Topic Word Selection for Topic Labelling. In: Proceedings of the 23rd international conference on computational linguistics. Beijing, China, pp 605–613
5. He D, Wang M, Khattak AM, Zhang L, Gao W (2019) Automatic labeling of topic models using graph-based ranking. IEEE Access 7:131593–131608
6. Alokaili A, Aletras N, Stevenson M (2019) Re-ranking words to improve interpretability of automatically generated topics. In: Proceedings of the 13th international conference on computational semantics—long Papers. Gothenburg, Sweden, pp 43–54
7. Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th international conference on computational semantics. Potsdam, Germany, pp 13–22
8. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing Semantic Coherence in Topic Models. In: Proceedings of the conference on empirical methods in natural language processing. Edingburgh, United Kingdom, pp 262–272
9. Rosner F, Hinneburg A, Röder M, Nettling M, Both A (2014) Evaluating topic coherence measures. In: Computing research repository (CoRR), pp 1–4. arXiv:1403.6397
10. Korencic D, Ristov S, Snajder J (2018) Document-based topic coherence measures for news media text. Expert Syst Appl 114:357–373
11. AlSumait L, Barbara D, Gentle J, Domeniconi C (2009) Topic Significance Ranking of LDA Generative Models. In: Proceedings of the European conference on machine learning and knowledge discovery in databases. Bled, Slovenia, pp 67–82
12. Kölbl L, Mühlroth C, Wiser F, Grottke M, Durst C (2019) Big data im Innovationsmanagement: Wie machine learning die Suche nach Trends und Technologien revolutioniert. HMD Praxis der Wirtschaftsinformatik 56(5):900–913
13. Thorleuchter D, Scheja T, Van den Poel D (2014) Semantic weak signal tracing. In: Expert systems with applications 41(11):5009–5016
14. Saritas O, Smith JE (2011) The big picture-trends, drivers, wild cards, discontinuities and weak signals. Futures 43(3):292–312
15. Mühlroth C, Grottke M (2018) A systematic literature review of mining weak signals and trends for corporate foresight. J Bus Econ 88(5):643–687
16. Lajoie EW, Bridges L (2014) Innovation decisions: using the Gartner hype cycle. Libr Leadersh Manage 28(4)
17. Lau JH, Baldwin T (2016) The sensitivity of topic coherence evaluation to topic cardinality. In: Proceedings of the North American chapter of the association for computational linguistics: human language technologies. San Diego, USA, pp 483–487

Laura Kölbl has a M.Sc. in statistics from the Ludwig-Maximilians-University Munich and is currently working at the chair of statistics and economics at the Friedrich-Alexander University Erlangen-Nuremberg. As part of a research project funded by the German Federal Ministry of Education and Research (BMBF), her research focuses on NLP and machine learning methods with a special interest for possible applications in the automatic detection of trends.

Michael Grottke received an M.A. in economics from Wayne State University, USA, as well as a Diploma degree in business administration, a Ph.D., and a Habilitation degree from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. He is now Principal Data Scientist at GfK SE, Nürnberg, Germany and Adjunct Professor at FAU. Before, he spent three years as a Research Associate and Assistant Research Professor in the Department of Electrical and Computer Engineering at Duke University, USA. His research focuses on stochastic modeling, statistical data analysis and machine learning for topics in information systems, computer science, and business administration, including crowdsourcing, social networks, software quality, and the automatic detection of upcoming technological trends. His research work has been funded by national and international organizations, such as the European Commission, the German Ministry of Education and Research, and NASA's Office of Safety and Mission Assurance. He has published more than 40 papers in international journals and peer-reviewed conference proceedings. He has acted as a Guest Editor for Special Editions of Performance Evaluation and the Journal of Systems and Software, and he is an Associate Editor of the International Journal of Performability Engineering and the IEEE Transactions on Reliability. He is a member of the IEEE, the German Statistical Society, and ESOMAR.

An Improved Ensemble Machine Learning Algorithm for Wearable Sensor Data Based Human Activity Recognition

Huu Du Nguyen, Kim Phuc Tran, Xianyi Zeng, Ludovic Koehl and Guillaume Tartare

Abstract Recent years have witnessed the rapid development of human activity recognition (HAR) based on wearable sensor data. One can find many practical applications in this area, especially in the field of health care. Many machine learning algorithms such as Decision Trees, Support Vector Machine, Naive Bayes, K-Nearest Neighbor, and Multilayer Perceptron are successfully used in HAR. Although these methods are fast and easy for implementation, they still have some limitations due to poor performance in a number of situations. In this chapter, we propose an improved machine learning method based on the ensemble algorithm to boost the performance of these machine learning methods for HAR.

1 Introduction

Human activity recognition (HAR) refers to a specific case of activity recognition that aims to recognize the actions of a person or to predict what a person is doing based on a series of observations on his/her actions and the environment. Thanks to the rapid development of advanced technologies, nowadays, the application of HAR

H. D. Nguyen
Division of Artificial Intelligence, Dong A University, Danang, Vietnam
e-mail: dunh@donga.edu.vn

K. P. Tran (✉) · X. Zeng
ENSAIT & GEMTEX, Roubaix, France
e-mail: kim-phuc.tran@ensait.fr

X. Zeng
e-mail: xianyi.zeng@ensait.fr

L. Koehl · G. Tartare
Ecole Nationale Supérieure des Arts et Industries Textiles, GEMTEX Laboratory,
59056 Roubaix, France
e-mail: ludovic.koehl@ensait.fr

G. Tartare
e-mail: guillaume.tartare@ensait.fr

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series
in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_13

becomes increasingly has become increasingly popular and more effective. The wide use of HAR can be seen in several practical fields such as security surveillance, human-computer interaction, military, and especially health care. For instance, the HAR is used for monitoring the activities of elderly people staying in rehabilitation centers for chronic disease management and disease prevention [1]. It is used to encourage physical exercises for assistive correction of children's sitting posture [2]. It is also applied in monitoring other behaviours such as abnormal conditions for cardiac patients and detection for early signs of illness [3]. A study on monitoring worker fatigue using wearable devices was conducted in Baghdadi et al. [4] where the authors focused on detecting changes in gait parameters. In the military, the exact information of locations, health conditions, and activities of the soldiers' are very useful for support them in decision making in both training and combat. More examples of applications of HAR can be seen in Ann and Theng [5].

There are two major methods using in HAR, involving external sensors-based method and wearable sensors-based method. Intelligent homes and cameras are two typical examples of external sensing. A number of limitations of this method have been discussed in Lara and Labrador [6], promoting the use of wearable sensors in HAR. Wearable sensor method refers to the use of smart electronic devices that are integrated into wearable objects or directly with the body in order to measure both biological and physiological sensor signals such as heart rate, blood pressure, body temperature, accelerometers, or other attributes of interest like motion and location. The use of a smart mobile phone containing several sensors such as gyroscope, camera, microphone, light, compass, accelerometer, proximity, and GPS can also be very effective for activity recognition [7]. However, the raw data from smart mobile phones are only effective for simple activities but not complex activities [8]. Thus, the extra sensors or sensing devices should be used for a better performance of the recognition.

In the literature, several machine learning algorithms have been suggested to deal with features extracted from raw signals to identify human activities. These machine learning methods are in general fast and easy to be implemented. However, they only bring satisfying results in a few scenarios because of relying heavily on the heuristic handcrafted feature extraction [9]. Recent years have witnessed the rapid development and application of deep learning, which has also achieved a remarkable efficiency in the HAR. Although the deep learning algorithms can automatically learn representative features due to its stacking structure, it has a disadvantage of requiring a large dataset for the training model. In fact, numerous data are available but it is sometimes difficult to access data that are labeled. It is also inappropriate for real-time human activity detection due to the high computation load [10]. The goal of this chapter is to improve the performance of these algorithms as well as keep the simplicity with implementation by applying an ensemble algorithm for machine learning based learners. In particular, we firstly conduct experiments with several machine learning classifiers such as Logistic Regression, Multilayer Perceptron, K-Nearest Neighbor, Support Vector Machine, and Random Forest. Then we apply a novel recognition model by using a voting algorithm to combine the performance of these algorithms for HAR. In fact, the ensemble algorithm has been suggested in

Catal et al. [11], leading to impressive results compared to other traditional machine learning methods. In this study, we enhance the performance of the method proposed in Catal et al. [11] by using more efficient classifiers as base models. The obtained results in this chapter have been presented in Nguyen et al. [12]

The chapter is organized as follows. In Sect. 2, we present a brief of the structure of a HAR system and related works. Section 3 describes the methodologies using in the study, including the sample generation process, the feature extraction, the basic machine learning algorithms, and the proposed voting classifier. The experimental results are shown in Sect. 4 and some concluding remarks are given in Sect. 5.

The following notations are used in this chapter—HAR: human activity recognition, LR: logistic regression; MLP: Multilayer Perceptron; KNN: K-nearest neighbour; SVM: suppor vector machine; RF: random forest.

2 A Brief Review of Human Activity Recognition Based on Wearable Sensor Data

In this section, we discuss briefly an overview of the HAR method based on wearable sensor data.

A large number of studies of wearable sensors-based HAR have been conducted in the literature. For an overview of this topic, Bulling et al. [13] gave a tutorial on HAR using body-worn inertial sensors. Shoaib et al. [14] described limitations and suggestions for online activity recognition using mobile phones. Lara and Labrador [6] made the problem more clearly by providing a generic data acquisition architecture for HAR systems and carrying out an extensive survey on wearable sensor-based HAR. Other surveys have been conducted in Stisen et al. [15] and Mukhopadhyay [16]. Figure 1 presents a graphic illustration of a general HAR system.

At the first step of the data collection process of a HAR system, wearable sensors are attached to the person's body to measure attributes of interest. These sensors are connected to an integration device such as a smart phone, personal digital assistance, a laptop or a customized embedded system to preprocess the data and send them to an application server for analysis. In general, the attributes measured using wearable sensors in a HAR system are composed of four main groups: environmental attributes, acceleration, location, and physiological signals. The first group aims to provide information pertaining to the surroundings of people like temperature, humidity, audio level, etc. [17]. The second one is to record most of the daily human basic activities like walking, running, lying, jumping, sitting, etc. [18, 19]. The location attribute is obtained from the global positioning system to provide information on the location of the subjects [20]. In the last group, several vital signs data related to the people's physical or biological health such as heart rate, respiration rate, skin temperature, skin conductivity are the main concerns Yin et al. [21]. Among these four groups of attributes, the sensors from the acceleration using triaxial accelerometers are perhaps the most broadly used to recognize ambulation activities.

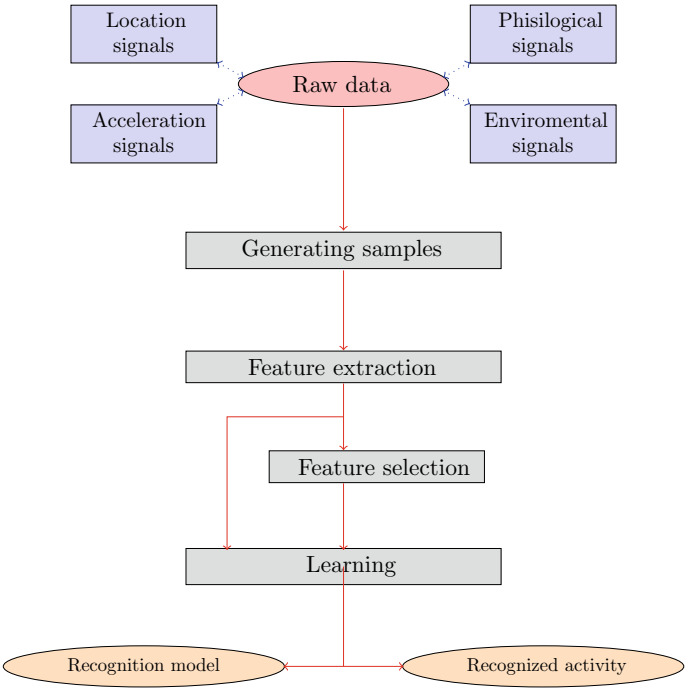


Fig. 1 A general HAR system based on wearable sensors

After obtaining raw signals from wearable devices, the next step is to generate the samples from these signals. In particular, the raw signals are divided into small parts of the same size, which are called temporal windows, and are used as training and test dataset to define the model. A number of techniques have been developed to separate the generated samples into training and testing sets such as k-fold cross-validation, leave-one-subject-out, hold-out and leave-one-sample-out. Among these methods, the leave-one-subject-out protocol can be considered as a special case of the cross-validation method by considering a subject as a fold and both of them are the traditional preferences. In the literature, the semi-non-overlapping-window (SNOW) technique is widely used to generate temporal windows from raw signals. According to this approach, the samples are firstly split into subparts or windows to create a set of data samples where each window is considered as an entire activity. This step is called a temporal sliding window and can be defined as

$$w = [s_{m-t}, \ldots, s_{m-1}, s_m]^T \tag{1}$$

where m denotes the current signal captured by the sensor and t is the temporal sliding window size [22]. Then, a validation protocol like leaveone-subject-out and 10-fold cross-validation considering an overlap of 50% between windows is applied

to produce sets for training and test from these data samples. Various values of t have been used in the literature, ranging from 0.08 s [18] to 12 s [23] or even up to 30 s [24]. It should be considered that if the value of t is too small, i.e. too short time windows, it may not provide sufficient information to fully describe the performed activity. In contrast, when one chooses a too large value of t, i.e. too large time windows, it might contain more than one activity [20].

Once the samples have been generated from raw signals, the next step is to apply feature extraction methods for filtering relevant information and obtaining quantitative measures from these raw data. This is because of the fact that the raw signals could be noisy and not representative of different activities. Many feature extraction approaches have been introduced for the purpose of activity recognition relying on different attributes of interest. For the case of acceleration signals, the statistical feature extraction, including both time- or frequency-domain features, is applied in most of the cases, for example in Baek et al. [25], Maurrer et al. [26], Ravi et al. [27], and Nyan et al. [28]. The discrete cosine transform method, the principal component analysis method and the autoregressive model coefficients method have also been considered to extract features from accelerometer signals by He and Jin [29, 30] with promising results. In Sani et al. [31], the feature extraction approaches for accelerometer data have been classified into handcrafted, frequency-transform and deep features where the deep features refer to the use of the convolutional neural networks method. Banos et al. [1] emphasized that feature extraction is considered to be the most critical stage in identifying the main unknown activity discriminant elements. Chen et al. [32] provided general formulas for a number of widely used features. Some of these formulas are presented for a given signal in the next section. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data was made in Preece et al. [33]. The method for extracting features from environmental attributes and the physiological attributes was discussed further in Lara and Labrador [6].

The feature selection is also a necessary step in the HAR structure, although, in some studies, this step can be skipped. Since several extracted features might contain redundant or irrelevant information that can have negative influence on the accuracy of the recognition process, feature selection can be applied to reduce computations and simplify the learning models. The forward-backward selection method was applied in Pirttikangas et al. [34]. Maurrer et al. [26] suggested applying the Correlation-based Feature Selection approach. The Minimum Redundancy and Maximum Relevance methods were utilized in Jatoba et al. [35]. More details on the technique of selecting feature can be seen in Banos et al. [1].

The last step in a HAR system is to use the classification algorithms to train the model. Due to its widely practical applications, a large number of machine learning algorithms have been suggested to deal with sensors data in the HAR such as Conditional Random Fields [36], Hidden Markov Models [38], Support Vector Machines [37], and K-Nearest Neighbor [39]. Other machine learning algorithms like J48 Decision Trees and Logistic Regression are also utilized for the HAR based on the accelerometer of a smartphone. An overview of the machine learning method using in HAR is provided in Lara and Labrador [6].

In general, when a few labeled data or certain knowledge is required, the machine learning pattern of recognition approaches can give satisfying results. However, there are several limitations of using these methods because of the following arguments: (1) these regular methods rely heavily on practitioners' experience with heuristic and handcrafted ways to extract interesting features; (2) the deep features are difficult to be learned; and (3) a large amount of well-labeled data to train the model is required [9]. This is the motivation for the use of deep learning in wearable sensor-based HAR recently. Hammerla et al. [40] used a 5-hidden-layer deep neural network to perform automatic feature learning and classification. A convolutional neural network has been studied in Chen and Xue [41], Graves et al. [42], Ha and Choi [43, 44], Jiang and Yin [45], Murad an Pyun [46], Yao et al. [47], and Hannink et al. [48]. A stacked autoencoder has been applied in Almaslukh et al. [49] and Wang et al. [50]. Sun et al. [51] proposed a hybrid deep framework based on convolution operations, long short-term memory recurrent units, and extreme learning machine classifiers. A comprehensive survey on deep learning for sensor-based activity recognition can be seen in Wang et al. [9].

Another method that is also widely used for HAR is ensemble learning. This method not only improves significantly the performance of traditional machine learning algorithms and but also avoid a requirement of a large dataset for training model of deep learning algorithms. Catal et al. [11] used a set of classifiers including J48 Decision Trees, Logistic Regression and Multilayer Perceptron, to recognize specific human activities like walking, jogging, sitting and standing based on accelerometer sensor of a mobile phone. Gandhi and Pandey [52] applied a hybrid ensemble classifier that combines the representative algorithms of Instance based learner, Naive Bayes Tree and Decision Tree algorithms using a voting methodology to 28 benchmark datasets and compared their method with other machine learning algorithms. A novel ensemble extreme learning machine for HAR using smartphone sensors has been presented in Chen et al. [10].

3 Methodology

In this section, we explain the method of generating samples as well as extracting features of data. The individual classifiers and an ensemble method using in this study are also presented.

3.1 Sample Generation Process for HAR

As discussed above, the technique is widely applied to generate temporal windows from raw signals. Using this method, however, leads to a significant disadvantage of being highly biased as pointed out in Jordao et al. [22]. That means part of the sample's content appears both in the training and testing at the same time. Then, the

authors proposed two new methods to handle this bias drawback, including full-non overlapping-window (FNOW) and leave-one-trial-out (LOTO) as follows.

- Full-Non-Overlapping-Window (FNOW): The reason for the bias problem of the SNOW process is because only an overlap of 50% between windows is considered. Therefore, in the FNOW process, the authors proposed using full non-overlapping windows to generate samples, which ensure no temporal overlapping in the windows. However, this leads to a reduced number of samples compared to the SNOW process since the temporal windows no longer overlap.
- Leave-One-Trial-Out (LOTO): This method allows us to avoid the drawback of both the SNOW method (producing biased results) and the FNOW method (generating a few samples). The activities from a trial are initially segmented and then 10-fold cross-validation is employed. The samples generated by the same trial are guaranteed not to appear simultaneously in both training and testing folds by using an incremental identifier assigned for a subject before performing the activities.

A figure illustrated the LOTO process for the example of 2-fold cross-validation is shown in Jordao et al. [22]. We apply this process in our study.

3.2 Feature Representation for HAR

Another important step in the process of HAR before applying machine learning algorithms is to extract features from raw data. In this study, we focus on common techniques used for acceleration signals which consist of time- and frequency-domain features. Typical time domain features are mean, standard deviation, variance, root squared mean, and percentiles; while typical frequency domain features include energy, spectral entropy, and dominant frequency. These measures are designed to capture the characteristics of the signal that are useful for distinguishing different classes of activities. Table 1 shows some widely used features from the literature.

Table 1 Hand-crafted features for both time and frequency domains

Time domain features	Frequency domain features
Mean	Dominant frequency
Standard deviation	Spectral centroid
Inter-quartile range	Energy
Kurtosis	Fast fourier transform
Percentiles	Discrete cosine transform
Mean absolute deviation	
Entropy	
Correlation between axes	

Following is the definition of some widely used features given a signal $Z = \{z_1, \ldots, z_n\}$ [6].

- Measures of center tendency like mean and root mean square (RMS):

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \tag{2}$$

$$RMS(Z) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} z_i^2}. \tag{3}$$

- Dispersion metrics like standard deviation (σ_z), variance (σ_z^2), and the mean absolute deviation (MAD)

$$\sigma_z = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2}, \tag{4}$$

$$MAD(Z) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}|z_i - \bar{z}|}. \tag{5}$$

- Measures of domain transform like the energy

$$Energy(Z) = \frac{1}{n}\sum_{i=1}^{n} F_i^2 \tag{6}$$

where F_i is the i-th component of the Fourier tranform of Z.

3.3 The Needed Concepts

In this section, we will present a brief of the needed machine learning algorithms for the ensemble algorithm. We denote $\mathbb{T} = \left\{(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \ldots, (\mathbf{x}_N, Y_N)\right\}$ the training data set from a d-dimension space, $\mathbf{x}_i \in \mathbb{R}^d$ and the corresponding label $Y_i \in \{c_1, \ldots, c_k\}$, where $c_1, \ldots, c_k$ are the k classes that are the targets of classification.

3.3.1 Logistic Regression

Logistic regression (LR) is a well-known statistical classification method for modeling a binary response variable, which takes only two possible values. It simply models probability of the default class. The LR method is commonly used because it is easy with implementation and it provides competitive results. Although LR is not

a classifier, it can still be used to make a classifier or prediction by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class and less than the cutoff as the other.

In the LR algorithm, the logarithmic of the odds ratio between classes is calculated as a linear function of the explanatory variable. A general form of LR is given by Friedman et al. [53]

$$P(Y = c_i \mid \mathbf{x}) = \frac{exp(\beta_{i0} + \beta_i^T \mathbf{x})}{1 + \sum_{j=1}^{k-1} exp(\beta_{j0} + \beta_j^T \mathbf{x})}, \; i = 1, \ldots, k-1,$$
$$P(Y = c_k \mid \mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{k-1} exp(\beta_{j0} + \beta_j^T \mathbf{x})}. \tag{7}$$

where $P(Y = c_i \mid \mathbf{x})$ is the probability of the response variable Y belonging to class c_i given the observation $\mathbf{x} \in \mathbb{R}^d$, and $\beta_{i0} \in \mathbb{R}$ and $\beta_i \in \mathbb{R}^d, i = 1, \ldots, k$ are the logistic regression parameters for each class. These parameters are estimated from the training data set $\mathbb{T}$. The class c of the output Y given the observations $\mathbf{x}^*$ is then defined by

$$c = \arg\max_{c_i} P(Y = c_i \mid \mathbf{x}^*), \; i = 1, \ldots, k. \tag{8}$$

More detail of the algorithm and its various applications can be seen in Hosmer et al. [54].

3.3.2 Multilayer Perceptron

A multilayer perceptron (MLP) is a classical type of feedforward artificial neural network. It is composed of more than one perceptron, which is an algorithm for learning a binary classifier. The MLP is able to overcome the principal weakness of the perceptron, that could only solve linearly separable problems.

At least three layers are included in a MLP with different functions: an input layer to receive the signal, an output layer to make a decision or a prediction about the input, and between these two layers are an arbitrary number of hidden layers to perform the true computational engine of the MLP. The nodes in MLP have fully connected in the sense that each node in one layer connects to every node in the following layer with a certain weight. Figure 2 illustrates a simple multilayer perceptron.

The mechanism within the MLP is as follows. Firstly, data are fed to the input layer. These data are then multiplied by the correponding weights of each node in the first layer of the hidden layers. The obtained results are passed through the node (or a neuron) using a nonlinear activation function, which is usually a sigmoid function. The output of each layer in the hidden layers will be the input for the adjacent one and the process similarly continues until the last layer in the hidden layers. Then, the output from this layer will be fed to the output layer before the final output is shown to make a decision.

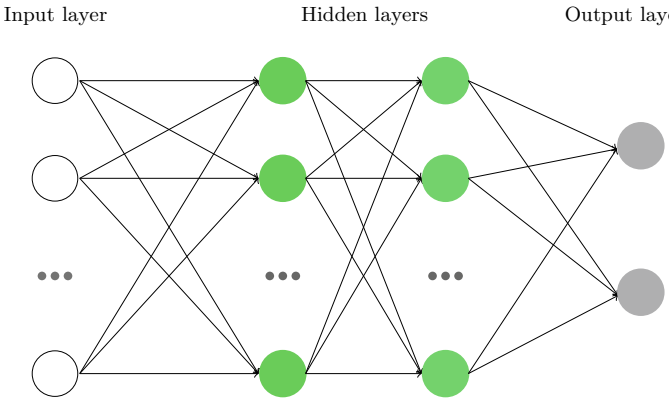


Fig. 2 A simple multilayer perceptron

In the first step of the traning process, the connection weights in each node of the hidden layers and the output layer are chosen randomly in the interval (0, 1). Then, these weights are trained to minimize the amount of error in the output compared to the expected result. The back-propagation learning technique is applied for the training of the network. The wide applications of MLP can be found in large fields such as speech recognition, image recognition, and machine translation software [55].

3.3.3 Support Vector Machine

Support Vector Machine (SVM) is a discriminative classifier belonging to supervised learning models. It constructs a hyperplane (a decision boundary that helps to classify the data points) or set of hyperplanes in a high or infinite dimensional space, which can be used for both classification and regression problem. The SVM is considered as one of the superior machine learning in HAR because of the prime generalization strength and highly accurate results.

Considering the context of binary classifcation, let $y_i \in \{-1, 1\}$ be the corresponding label of $\mathbf{x}_i$, $i = 1, \ldots, N$. The SVM algorithm learns from the training data a function $f(x) = w^T\mathbf{x} + b$, where $w \in \mathbb{R}^d$, such that $f(\mathbf{x}_i) > 0$ for every $y_i = 1$ and $f(\mathbf{x}_i) < 0$ for $y_i = -1$. In practice, there are many possible hyperplanes that could separate the two classes of data points. The SVM aims to find a hyperplane that has maximum margin, i.e. the maximum distance between data points of both classes. In a general case, when the obsesrvations are non-separable, the algorithm introduces non-negative slack variables ξ_i to allow for misclassifcation of some observation $\mathbf{x}_i$, $i = 1, \ldots, N$. $\xi_i = 0$ means that $\mathbf{x}_i$ is on the correct side of the separating hyperplane. Meanwhile, the value $\xi_i > 0$ indicates the magnitude that $\mathbf{x}_i$ violates the soft margin constraints. Figure 3 illustrates the scenario. The learning process in SVM involves estimating the parameters w and b such that they satisfy conditions on the training data:

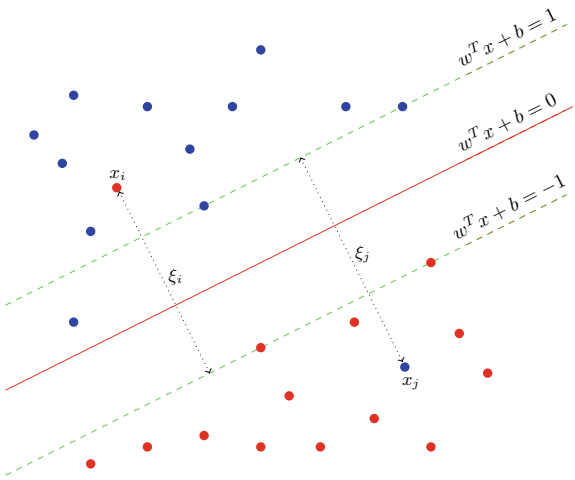


Fig. 3 Support Vector Machine with slack variables

$$\begin{aligned} w^T \mathbf{x}_i + b &\geqslant 1 - \xi_i \text{ for } y_i = 1, \\ w^T \mathbf{x}_i + b &\leqslant -1 + \xi_i \text{ for } y_i = -1, \\ \xi_i &\geqslant 0 \; \forall i = 1, \ldots, N. \end{aligned} \tag{9}$$

The overall goal for SVM is to minimize the number of misclassifcations. That is, the SVM training is equivalent to solving a linear constrained quadratic programming problem

$$\arg\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{10}$$

subject to $y_i(w^T \mathbf{x}_i + b) \geqslant 1 - \xi_i \; \forall \mathbf{x}_i \in \mathbb{T}$ and $\xi_i \geqslant 0, i = 1, \ldots, N$.

This algorithm can be applied to various problems such as text and hypertext categorization [56], images classification [57], and hand-written characters recognition [58]. A review on rule extraction and the main features of algorithm from SVM are presented in Barakat and Bradley [59].

3.3.4 Gaussian Naive Bayes

Gaussian Naive Bayes refers to a naive Bayes classifier, which is a simple probabilistic classifier based on applying Bayes' theorem with *naive* independence assumptions, dealing with continuous data in the case that the continuous values associated with each class are distributed following a Gaussian distribution. The output Y is classified into a class c given the observations $\{\mathbf{x}_1^*, \mathbf{x}_2, \ldots, \mathbf{x}_n^*\}$ if c satisfies the following condition:

$$c = \arg\max_{c_i} P(c_i) \prod_{j=1}^{n} \varphi(\mathbf{x}_j^* \mid c_i),\ i = 1, \ldots, k, \tag{11}$$

where $\varphi(. \mid c_i)$ denotes a probability density function of the multivariate normal distribution of the class c_i. From the Bayesian statistic point of view, c represents a maximum a posteriori estimate. The model is trained by estimating the parameters of $\varphi(. \mid c_i)$ in each class from the training dataset $\mathbb{T}$.

The major advantage of the naive Bayes classifier is its simplicity and short computational time for training. It can also often outperform some other more sophisticated classification methods [60]. The attempts to overcome the independence assumption using in this algorithm has been discussed in Rennie et al. [61] and Kotsiantis et al. [62].

3.3.5 K-Nearest Neighbor

The K-Nearest Neighbor (KNN) belongs to a class of non-parametric algorithms known as prototype methods. It is considered as an instance-based since a new object is classified based on the distance from its K neighbors in the training set and the corresponding weights assigned to the contributions of the neighbors. The distance commonly used for continuous variables in the algorithm is the Euclidean distance. The weights are defined based on a simple principle, that is similar things are near to each other and the nearer the neighbors are, the more they contribute to the average. A feature of the KNN algorithm is that it requires no model to be fit. An illustration of the algorithm can be seen in Fig. 4. The new observation $\mathbf{x}^*$ is grouped into the class c based on the majority class among the K nearest neighbors:

$$c = \arg\max_{c_i} \sum_{(\mathbf{x}_i, Y_i) \in \mathrm{NN}(\mathbf{x}^*, \mathbb{T}, K)} \mathbb{1}(Y_i = c_i),\ i = 1, \ldots, k, \tag{12}$$

where $\mathrm{NN}(\mathbf{x}^*, \mathbb{T}, K)$ denotes the K nearest neighbors of $\mathbf{x}^*$ in $\mathbb{T}$. As an example, in Fig. 4, the red star is predicted to be in the class of blue circle points if $K = 3$. In contrast, when $K = 5$, it is considered belonging to the class of green rectangle points.

The parameter K has a significant effect on the performance of the KNN algorithm. It should be considered that there is no principled way to choose K, except through cross-validation or similar, computationally-expensive technique Guo et al. [63]. The small values of K may lead to overfitting, less stable or influenced by noise problems. In contrast, the large values of K can make the classification less precise or higher bias. Kotsiantis et al. [62] discussed some practical reasons for the choice of K. A review of data classification using this KNN algorithm is given in Kataria and Singh [64].

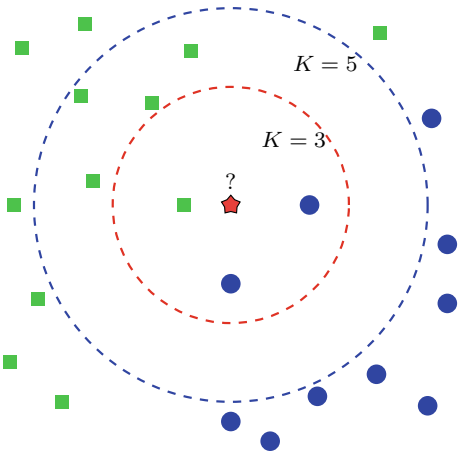


Fig. 4 K-nearest neighbor

3.3.6 Random Forest

Random forest is an ensemble learning method that allows to classify objects by voting the results of individual decision trees. In the traning process, each tree in a random forest learns from a bootstrap sample of the training data. The samples are drawn with replacement, i.e, some samples can be used several times in an individual tree. Unlike the bagging algrithm, not all the features of the bootstrapped data are used to grow the tree. Insread, only a subset of features are considered to split nodes in the tree. This makes the RF algorithm improve the correlation among the trees in the bagging algorithm. In particular, a RF is described as follows.

Input: The traning dataset $\mathbb{T} = \left\{(\mathbf{x}_i, Y_i)_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^d, Y_i \in \{c_1, \ldots, c_k\}\right\}$,
The number of trees K,
The size of subspace *mtry* where $m < d$.
Output: A random forest for classification a new object.
Do
1. For $i = 1$ to K:
   - Draw a bootstrap sample $\mathbb{T}_i$ from $\mathbb{T}$.
   - Grow a random-forest tree T_i from $\mathbb{T}_i$, by repeating the three following steps:
     (*i*) Select randomly m features.
     (*ii*) Choose the best feature among the m that minimizes the impurity.
     (*iii*) Split the node into two children nodes.
2. Combine the trees $\{T_i\}_{i=1}^K$ to form a random forest.

The final classification $\hat{T}(\mathbf{x}^*)$ given the observation $\mathbf{x}^*$ is defined by

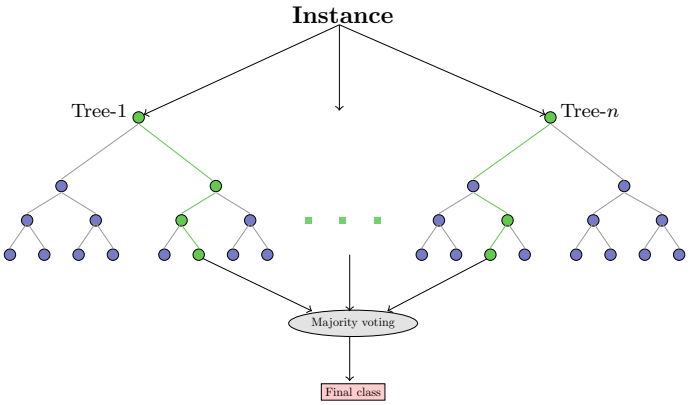


Fig. 5 An illustration of the random forest

$$\hat{T}(\mathbf{x}^*) = \text{majority vote } \left\{T_i(\mathbf{x}^*)\right\}, \tag{13}$$

where $T_i(\mathbf{x}^*)$ denotes the class of $\mathbf{x}^*$ according to the tree T_i from the RF algorithm (Fig. 5).

A method of building a forest of uncorrelated trees using a CART (Classification And Regression Tree) and several ingredients forming the basis of the modern practice of random forests have been introduced in Breiman [65].

3.4 Proposed Voting Classifier

Each machine learning algorithm for the classification presented above has their own advantages and disadvantages. An ensemble method is then proposed to combines these algorithms in order to obtain an optimal classifiable model.

Suppose that we have R classifiers where each of them represent a given object $\mathbf{x}^*$ by a distinct measurement vector, say $\mathbf{x}_i^*$ for the ith classifier, $i = 1 \ldots, R$. A principle in the ensemble learning is that the new object $\mathbf{x}^*$ is classified into the group c_i if

$$P(c_i \mid \mathbf{x}_1^*, \ldots, \mathbf{x}_R^*) = \max_j P(c_j \mid \mathbf{x}_1^*, \ldots, \mathbf{x}_R^*). \tag{14}$$

Several ensemble rules for combining the multiple classification results of different classifiers, including Min rule, Maximum rule, Median rule, and Majority vote rule have been proposed in Kittler et al. [66]. In this study, we apply the voting rule, which is a simple but powerful technique, for combining the aforementioned algorithms. By this strategy, let Δ_{ij} denote a binary-valued function where

$$\Delta_{ij} = \begin{cases} 1 & \text{if } P(c_i|\mathbf{x}_j^*) = \max_t P(c_t|\mathbf{x}_j^*),\ t = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

then $\mathbf{x}^*$ will be assigned to the class c_i if the following condition is satisfied:

$$\sum_{j=1}^{R} \Delta_{ij} = \max_t \sum_{j=1}^{R} \Delta_{tj},\ t = 1, \ldots, k. \tag{16}$$

Detail of three versions of voting, involving unanimous voting, majority voting and plurality voting, can be found in Wolpert [67]. In unanimous voting, the final decision is approved by all the base learners. In majority voting, more than 50% vote is required for final decision. In plurality voting, most of the votes decides the final result, i.e. the activity which recived the most votes from component algorithms is decided.

4 Experimental Results

In this section, we show the experimental results of our proposed method and compare these results with the ones obtained by using the method suggested in Catal et al. [11]. The two following data sets are considered:

- MHEALTH: This dataset is based on a new framework for agile development of mobile health applications suggested by Banos et al. [68]. It consists of body motion and vital signs recordings for ten subjects of diverse profile while performing several physical activities, involving standing still,sitting and relaxing, lying down, walking, climbing stairs, waist bends forward, frontal elevation of arms, knees bending, cycling, jogging, running, and jump front and back. The sensor positioned on the subject's chest provides accelerometer signals, gyroscope signals, magnetometer signals, and electrocardiogram signals. All sensing modalities are recorded at a sampling rate of 50 Hz, which is considered sufficient for capturing human activity. However, since the electrocardiogram signal is damaged in a large number of samples as discussed in Jordao et al. [22], we consider only the first three signals.
- USC-HAD: The USC-HAD, standing for the University of Southern California Human Activity Dataset, is introduced in Zhang and Sawchuk [69]. This is specifically designed to include the most basic and common human activities in daily life, involving walking forward, walking left, walking right, walking upstairs, walking downstairs, running forward, jumping, sitting, standing, sleeping, elevator up, and elevator down. The data are collected from 14 subjects (7 male, 7 female) which consider the diversity of the subjects enrolled from four main factors: gender, age, height, and weight. A description for each activity can be seen Zhang and Sawchuk [69] while the detail of the data set is available at the website of the authors.

Table 2 Main fearutres of two datasets. A: Accelerometer, G: Gyroscope, M: Magnetometer

Feature	Data	
	MHELTH	USCHAD
Frenquency (Hz)	50	100
Number of sensors	3 (A,G, M)	2 (A, G)
Number of activities	12	12
Number of subjects	10	14
Number of trials	262	840
Number of samples	2555	9824

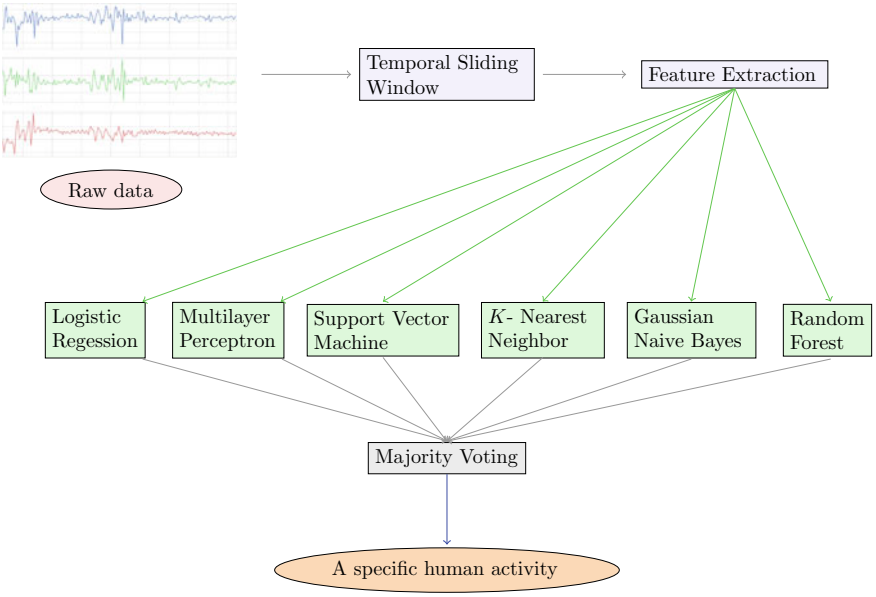


Fig. 6 The proposed framework for HAR using an ensemble algorithm

Table 2 presents a brief summary of the main features of the two data sets.

The first step in the experimental setup is temporal sliding window where the samples are split into subwindows and each subwindow is considered as an entire activity. We apply the same value of the temporal sliding window size which is $t = 5$ s as suggested in Jordao et al. [22]. Then, the important features from these raw signals (presented in Table 1) are extracted to filter relevant information and to give the input for classifiers. The output of ensemble algorithm is a recognized specific human activity. Figure 6 presents the proposed framework for HAR using ensmble method.

Based on the MHEALTH and USC-HAD data set, twelve basic and common activities in people's daily lives is the target to be recognized. The performance of

proposed method is evaluated through *Accuracy*, *Recall* and *F-score* metrics. These measures are defined based on the following values:

- *TP* (True Positive): the number of samples recognized correctly as activities,
- *TN* (True Negative): the number of samples correctly recognized as not activities,
- *FP* (False Positive): the number of samples incorrectly recognized as activities,
- *FN* (False Negative): the number of samples incorrectly diagnosed as not activities.

In particular, the accuracy metric, the recall metric and the F-score metric are defined as:

- $\text{Accuracy} = \dfrac{TP + TN}{TP + FP + TN + FN}$,
- $\text{Recall} = \dfrac{TP}{TP + FN}$,
- $\text{F-score} = 2 \times \dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

where $\text{Precision} = TP/(TP + FP)$.

Our computation was performed on a platform with 2.6 GHz Intel(R) Core(TM) i7 and 32GB of RAM. We perform the experience on the two datasets mentioned above using both our proposed method and the method suggested in Catal et al. [11]. The experimental results are presented in Table 3. It can be seen that our method leads to higher values of all the measures being considered. For example, for the MHEALTH data set, our method leads to $Accuracy = 94.72\%$ compared to $Accuracy = 93.87\%$ from the Catal et al.'s method. Similarly, for the USCHAD data set, *Recall* is equal to 83.20% corresponding to our proposed method, which is relatively larger than the value $Recall = 81.74\%$ corresponding to the Catal et al.'s method. The confident intervals obtained from our experiment are also narrower than those obtained by Catal et al. [11] for both datasets. For example, with the MHELTH dataset, the interval of F-score of our proposed method is (0.9100, 0.9725) compared to (0.8841, 0.9838) of the one from Catal et al.'s method. That is to say, the proposed method in this study outperforms the method used in Catal et al. [11]. Moreover, the obtained results show that the performance of an ensemble learning algorithm can be significantly improved by combining better machine learning algorithms. This should be considered in design a new ensemble method for HAR.

5 Conclusion and Future Work

Activity recognition increasingly plays an important role in many practical applications, especially in health care. Increasing the performance of recognition algorithms is a major concern for researchers. In this chapter, we have proposed an improved machine learning method for wearable sensor data based HAR using the ensemble

Table 3 The obtained results and a comparison with the method proposed by Catal et al. [11]

Measures	MHELTH		USCHAD	
	Proposed method	Catal et al.'s method	Proposed method	Catal et al.'s method
Accuracy	0.9472 (*0.9191, 0.9753*)	0.9387 (*0.8941, 0.9834*)	0.8690 (*0.8528, 0.8852*)	0.8528 (*0.8335, 0.8721*)
Recall	0.9498 (*0.9240, 0.9756*)	0.9423 (*0.9006, 0.9840*)	0.8320 (*0.8152, 0.8488*)	0.8174 (*0.7978, 0.8369*)
F-score	0.9412 (*0.9100, 0.9725*)	0.9339 (*0.8841, 0.9838*)	0.8190 (*0.8027, 0.8354*)	0.8160 (*0.7961, 0.8358*)

algorithm. By combining better classifiers, our proposed method improves significantly the previous study [11] in all measures of comparison.

In the future, we would like to address the problem of wearable sensor data based HAR using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The motivation for using these methods is because the CNN offers advantages in selecting good features and the LSTM has been proven to be good at learning sequential data.

References

1. Banos O, Damas M, Pomares H, Prieto A, Rojas I (2012) Daily living activity recognition based on statistical feature quality group selection. Expert Syst Appl 39(9):8013–8021
2. Chang Y, Chen S, Huang J (2011) A kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. Res Dev Disabil 32(6):2566–2570
3. Kańtoch E, Augustyniak P (2012) Human activity surveillance based on wearable body sensor network. In: Computing in cardiology. IEEE, pp 325–328
4. Baghdadi A, Cavuoto LA, Jones-Farmer A, Rigdon SE, Esfahani ET, Megahed FM (2019) Monitoring worker fatigue using wearable devices: a case study to detect changes in gait parameters. J Qual Technol. https://doi.org/10.1080/00224065.2019.1640097
5. Ann OC, Theng LB (2014) Human activity recognition: a review. In: IEEE international conference on control system, computing and engineering (ICCSCE). IEEE, pp 389–393
6. Lara OD, Labrador MA (2013) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tuts 15(3):1192–1209
7. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. SIGKDD Explor 12(2):74–82
8. Dernbach S, Das B, Krishnan N, Thomas BL, Cook DJ (2012) Simple and complex activity recognition through smart phones. In: 2012 eighth international conference on intelligent environments. IEEE, pp 214–221
9. Wang J, Chen Y, Hao S, Peng X, Hu L (2018) Deep learning for sensor-based activity recognition: a survey. Pattern Recogn Lett 119:3–11
10. Chen Z, Jiang C, Xie L (2018) A novel ensemble elm for human activity recognition using smartphone sensors. IEEE Trans Ind Inform 15(5):2691–2699
11. Catal C, Tufekci S, Pirmit E, Kocabag G (2015) On the use of ensemble of classifiers for accelerometer-based activity recognition. Appl Soft Comput 37:1018–1022

12. Nguyen HD, Tran KP, Zeng X, Koehl L, Tartare G (2019) Wearable sensor data based human activity recognition using machine learning: a new approach. In: ISSAT international conference on data science in business, finance and industry
13. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput Surv 46(3):33
14. Shoaib M, Bosch S, Incel O, Scholten H, Havinga P (2015) A survey of online activity recognition using mobile phones. Sensors 15(1):2059–2085
15. Stisen A, Blunck H, Bhattacharya S, Prentow TS, Kjærgaard MB, Dey A, Sonne T, Jensen M (2015) Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. In: ACM conference on embedded networked sensor systems. ACM, pp 127–140
16. Mukhopadhyay S (2015) Wearable sensors for human activity monitoring: a review. IEEE Sens J 15(3):1321–1330
17. Parkka J, Ermes M, Korpipaa P, Mantyjarvi J, Peltola J, Korhonen I (2006) Activity classification using realistic data from wearable sensors. IEEE Trans Inf Technol Biomed 10(1):119–128
18. Berchtold M, Budde M, Schmidtke HR, Beigl M (2010) An extensible modular recognition concept that makes activity recognition practical. In: Annual conference on artificial intelligence. Springer, pp 400–409
19. Brezmes T, Gorricho J, Cotrina J (2009) Activity recognition from accelerometer data on a mobile phone. In: International work-conference on artificial neural networks. Springer, pp 796–799
20. Reddy S, Mun M, Burke J, Estrin D, Hansen M, Srivastava M (2010) Using mobile phones to determine transportation modes. ACM Trans Sens Netw 6(2):13
21. Yin J, Yang Q, Pan JJ (2008) Sensor-based abnormal human-activity detection. IEEE Trans Knowl Data Eng 20(8):1082–1090
22. Jordao A, Nazare AC, Sena J, Robson Schwartz W (2018) Human activity recognition based on wearable sensor data: a standardization of the state-of-the-art. arXiv:1806.05226
23. Lara OD, Pérez AJ, Labrador MA, Posada JD (2012) Centinela: a human activity recognition system based on acceleration and vital sign data. Pervasive Mob Comput 8(5):717–729
24. Tapia EM, Intille SS, Haskell W, Larson K, Wright J, King A, Friedman R (2007) Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor. In: International symposium on wearable computers. IEEE, pp 37-40
25. Baek J, Lee G, Park W, Yun B (2004) Accelerometer signal processing for user activity detection. In: International conference on knowledge-based and intelligent information and engineering systems. Springer, pp 610–617
26. Maurer U, Smailagic A, Siewiorek DP, Deisher M (2006) Activity recognition and monitoring using multiple sensors on different body positions. In: International workshop on wearable and implantable body sensor networks, pp 112–116
27. Ravi N, Dandekar N, Mysore P, Littman ML (2005) Activity recognition from accelerometer data. AAAI 5:1541–1546
28. Nyan M, Tay F, Seah K, Sitoh Y (2006) Classification of gait patterns in the time-frequency domain. J Biomech 39(14):2647–2656
29. He Z, Jin L (2008) Activity recognition from acceleration data using ar model representation and SVM. In: 2008 international conference on machine learning and cybernetics. IEEE, pp 2245–2250
30. He Z, Jin L (2009) Activity recognition from acceleration data based on discrete consine transform and SVM. In: 2009 IEEE international conference on systems, man and cybernetics. IEEE, pp 5041–5044
31. Sani S, Wiratunga N, Massie S, Cooper K (2016) Selfback activity recognition for self-management of low back pain. In: International conference on innovative techniques and applications of artificial intelligence. Springer, pp 281–294
32. Chen Y, Yang J, Liou S, Lee G, Wang J (2008) Online classifier construction algorithm for human activity detection using a tri-axial accelerometer. Appl Math Comput 205(2):849–860

33. Preece SJ, Goulermas JY, Kenney L, Howard D (2008) A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Trans Biomed Eng 56(3):871–879
34. Pirttikangas S, Fujinami K, Nakajima T (2006) Feature selection and activity recognition from wearable sensors. In: International symposium on ubiquitious computing systems. Springer, pp 516–527
35. Jatoba LC, Grossmann U, Kunze C, Ottenbacher J, Stork W (2008) Context-aware mobile health monitoring: evaluation of different pattern recognition methods for classification of physical activity. In: International conference of the engineering in medicine and biology society. IEEE, pp 5250–5253
36. Vail DL, Veloso MM, Lafferty JD (2007) Conditional random fields for activity recognition. In: International joint conference on autonomous agents and multiagent systems. ACM, p 235
37. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2012) Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: International workshop on ambient assisted living. Springer, pp 216–223
38. Lee Y, Cho S (2011) Activity recognition using hierarchical hidden markov models on a smartphone with 3D accelerometer. In: International conference on hybrid artificial intelligence systems. Springer, pp 460–467
39. Ayu MA, Ismail SA, Matin AFA, Mantoro T (2012) A comparison study of classifier algorithms for mobile-phone's accelerometer based activity recognition. Procedia Eng 41:224–229
40. Hammerla NY, Halloran S, Ploetz T (2016) Deep, convolutional, and recurrent models for human activity recognition using wearables. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence. Newcastle University
41. Chen Y, Xue Y (2015) A deep learning approach to human activity recognition based on single accelerometer. In: 2015 IEEE international conference on systems, man, and cybernetics (MSC). IEEE, pp 1488–1492
42. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: International conference on acoustics, speech and signal processing. IEEE, pp 6645–6649
43. Ha S, Choi S (2016) Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: International joint conference on neural networks (IJCNN). IEEE, pp 381–388
44. Ha S, Yun J, Choi S (2015) Multi-modal convolutional neural networks for activity recognition. In: International joint conference on systems, man, and cybernetics. IEEE, pp 3017–3022
45. Jiang W, Yin Z (2015) Human activity recognition using wearable sensors by deep convolutional neural networks. In: International conference on multimedia. ACM, pp 1307–1310
46. Murad A, Pyun J (2017) Deep recurrent neural networks for human activity recognition. Sensors 17(11):2556
47. Yao S, Hu S, Zhao Y, Zhang A, Abdelzaher T (2017) Deepsense: a unified deep learning framework for time-series mobile sensing data processing. In: International conference on world wide web. International World Wide Web Conferences Steering Committee, pp 351–360
48. Hannink J, Kautz T, Pasluosta CF, Gaßmann K, Klucken J, Eskofier BM (2017) Sensor-based gait parameter extraction with deep convolutional neural networks. IEEE J Biomed Health 21(1):85–93
49. Almaslukh B, AlMuhtadi J, Artoli A (2017) An effective deep autoencoder approach for online smartphone-based human activity recognition. Int J Comput Sci Netw 17(4):160–165
50. Wang A, Chen G, Shang C, Zhang C, Liu L (2016) Human activity recognition in a smart home environment with stacked denoising autoencoders. In: International conference on web-age Information management. Springer, pp 29–40
51. Sun J, Fu Y, Li S, He J, Xu C, Tan L (2018) Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors. J Sens
52. Gandhi I, Pandey M (2015) Hybrid ensemble of classifiers using voting. In: International conference on green computing and Internet of Things. IEEE, pp 399–404

53. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in Statistics, New York (2001)
54. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, vol 398. Wiley
55. Wasserman PD, Schwartz T (1988) Neural networks. II. what are they and why is everybody so interested in them now? IEEE Expert 3(1):10–15
56. Pradhan S, Ward WH, Hacioglu K, Martin JH, Jurafsky D (2004) Shallow semantic parsing using support vector machine. In: The human language technology conference of the North American chapter of the association for computational linguistics
57. Barghout L (2015) Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. In: Granular computing and decision-making. Springer, pp 285–318
58. Decoste D, Schölkopf B (2002) Training invariant support vector machines. Mach Learn 46(1–3):161–190
59. Barakat N, Bradley AP (2010) Rule extraction from support vector machines: a review. Neurocomputing 74(1–3):178–190
60. Domingos P, Pazzani M (1997) On the optimality of the simple bayesian classifier under zero-one loss. Mach Learn 29(2–3):103–130
61. Rennie JD, Shih L, Teevan J, Karger D (2003) Tackling the poor assumptions of Naive Bayes text classifiers. In: International conference on machine learning, pp 616–623
62. Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
63. Guo G, Wang H, Bell D, Bi Y, Greer K (2003) KNN model-based approach in classification. In: OTM confederated international conferences on the move to meaningful internet systems. Springer, pp 986–996
64. Kataria A, Singh M (2013) A review of data classification using k-nearest neighbour algorithm. Int J Emerg Technol Adv Eng 3(6):354–360
65. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
66. Kittler J, Hater M, Duin R (1996) Combining classifiers. In: International conference on pattern recognition. IEEE, pp 897–901
67. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259
68. Banos O, Garcia R, Holgado-Terriza JA, Damas M, Pomares H, Rojas I, Saez A, Villalonga C (2014) Mhealthdroid: a novel framework for agile development of mobile health applications. In: International workshop on ambient assisted living. Springer, pp 91–98
69. Zhang M, Sawchuk AA (2012) USC-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: ACM conference on ubiquitous computing. ACM, pp 1036–1043

Huu Du Nguyen is a member of the Division of Artificial Intelligence, Dong A University, Danang, Vietnam. He is also a member of the Faculty of Information Technology, Vietnam National University of Agriculture, Hanoi, Vietnam. His research is related to reliability assessment, statistical process control, machine learning and deep learning.

Kim Phuc Tran obtained his Ph.D. degree in Automation and Applied Informatics at the Université de Nantes, Nantes, France, in 2016. He is currently an Associate Professor in Automation and Industrial Informatics at ENSAIT and GEMTEX, Roubaix, France. His research is focused on real-time anomaly detection techniques for Big Data, data mining, and machine learning techniques.

Xianyi Zeng received his Ph.D. degree from the Centre d'Automatique, Université des Sciences et Technologies de Lille, France, in 1992. He is currently a Full Professor of ENSAIT, Roubaix, France, and director of the GEMTEX National Laboratory. He is also an Associate Editor of *International Journal of Computational Intelligence System and Journal of Fiber Bioengineering and*

Informatics, a Guest Editor of Special Issues for six international journals, and a Senior Member of IEEE. He has organized 12 international conferences and workshops since 1996. Since 2000, he has been the leader of three European projects, four national research projects funded by the French government, three bilateral research cooperation projects, and more than 20 industrial projects. His research interests include: (1) intelligent decision support systems for fashion and material design, (2) modeling of human perception and cognition on industrial products, (3) intelligent wearable systems.

Ludovic Koehl was born in Paris, France, on August 12th, 1969. He is graduated from the "Ecole Nationale Superieure des Arts et Industries Textiles (ENSAIT)", Textiles Engineering Institution (France) in 1994 and received his Ph.D. degree in automation from the University of Lille, northern France, in 1998. Since 2006, he is qualified for leading research activities at national level. Since 2010, he works as full Professor in the ENSAIT Textile Engineering Institute, Roubaix, France and was the Deputy Director and the Director of the GEMTEX Lab., Roubaix France. To date he is the scientific advisor of the research group 'Human Centered Design' in GEMTEX research lab. His research interests include pattern recognition, data mining, computer modeling, big data management, recommendation systems, traceability and their applications in textile industry. Since 1998, he published more than 92 papersand took part of numerous research projects.

Guillaume Tartare got a Ph.D. degree in automatic, computer science and signal and image processing at LISIC (computer science laboratory) and INSERM (French institute of health and medical research) in 2014. In his thesis, he purposed a number of methods for detecting and classifying DCE MRI signals (Dynamic Contrast-Enhanced Magnetic Resonance Imaging). He is a member of SFGBM (French society of medical science). He is currently a associate Professor with the French Engineer School, Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT), Roubaix, France. Now his research interests include physiological signal acquisition and processing, design of intelligent garments by integrating sensors and a decision support system and morphology analysis.

Average Failure Rate and Its Applications of Preventive Replacement Policies

Xufeng Zhao, Jiajia Cai, Satoshi Mizutani and Toshio Nakagawa

Abstract When a mission arrives at a random time and lasts for an interval, it becomes an important constraint to plan preventive replacement policies, as the unit should provide reliability and no maintenance can be done during the mission interval. From this viewpoint, this chapter firstly gives a definition of an average failure rate, which is based on the conditional failure probability and the mean time to failure, given that the unit is still survival at the mission arrival time. Next, age replacement models are discussed analytically to show that how the average failure rate function appears in the models. In addition, periodic replacement models with minimal repairs are discussed in similar ways. Numerical examples are given when the mission arrival time follows a gamma distribution and the failure time of the unit has a Weibull distribution.

Keywords Age replacement · Minimal repair · Failure rate · Mission interval · Reliability

1 Introduction

Preventive replacement policies have been studied extensively in literatures [1–7]. Barlow and Proschan [1] have firstly given an age replacement model for a finite operating time span, where the unit operates from installation to a fixed interval caused by external factors, and it is replaced at the end of the interval even if no

The chapter submitted to H. Pham (Ed), Reliability and Statistical Computing, Springer.

X. Zhao (✉) · J. Cai
College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
e-mail: reliab@outlook.com

S. Mizutani · T. Nakagawa
Department of Business Administration, Aichi Institute of Technology, Toyota 470-0392, Japan

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_14

failure has occurred. When the finite time span becomes a random interval with working cycles, during which, it is impossible to perform maintenance policies, optimal replacement policies with random works have been discussed [3, 6, 8, 9].

When the working cycles are taken into account for planning replacement polices, Zhao and Nakagawa [10] proposed the policies of replacement first and replacement last, that would become alternatives in points of cost rate, reliability and maintainability. Replacement first means the unit is replaced preventively at events such as operating time, number of repairs, or mission numbers, etc, whichever takes place first, while replacement last means the unit is replaced preventively at the above events, whichever takes place last. It has been shown that replacement last could let the unit operate working cycles as longer as possible while replacement first are more easier to save total maintenance cost [10]. More recent models of replacement first and replacement last can be found in [11–15].

In this chapter, the above working cycle is reconsidered as mission interval, and we suppose that the arrival time of a mission is a random variable rather than it begins from installation and lasts for an interval, during which, the unit should provide reliability and no maintenance can be done. The typical example is maintaining a hot spare for a key unit in a working system, in which, the spare unit should be active at the time when the key unit fails and provide system reliability for an interval when the key unit is unavailable. From this viewpoint, this chapter discusses preventive replacement policies for random arrival of missions. For this, an average failure rate is firstly given based on the conditional failure probability and the mean time to failure, given that the unit is still survival at time t. We next formulate and optimize the models of age replacement policies and the periodic policies with minimal repairs in analytical ways. Numerical examples are given when the mission arrival time follows a gamma distribution and the failure time of the unit has a Weibull distribution.

2 Average Failure Rate

It is assumed that a unit has a general failure distribution $F(t) \equiv \Pr\{X \le t\}$ with a density function $f(t) \equiv \mathrm{d}F(t)/\mathrm{d}t$ and a finite mean $\mu \equiv \int_0^\infty \overline{F}(t)\mathrm{d}t$. The conditional failure probability is given by [2]:

$$\lambda(t; x) \equiv \frac{F(t+x) - F(t)}{\overline{F}(t)} \quad (0 < x < \infty), \tag{1}$$

which represents the probability that the unit fails in interval $[t, t+x]$, given that it is still survival at time t. Note that $0 \le \lambda(t; x) \le 1$. When $x \to 0$, $\lambda(t; x)/x$ becomes an instant failure rate:

$$h(t) \equiv \frac{f(t)}{\overline{F}(t)} = -\frac{1}{\overline{F}(t)} \frac{\mathrm{d}\overline{F}(t)}{\mathrm{d}t}. \tag{2}$$

We usually suppose, in modeling maintenance policies, that $h(t)$ increases with t from $h(0) = 0$ to $h(\infty) \equiv \lim_{t\to\infty} h(t)$ that might be infinity, i.e., $\lambda(t; x)$ increases with t from $F(x)$ to 1.

We next define:

$$F(t; x) = \frac{\int_t^{t+x} \overline{F}(u)\mathrm{d}u}{\overline{F}(t)}, \tag{3}$$

which means the mean time to failure, given that the unit is still survival at time t. Obviously, when $t \to 0$, $F(t; x)$ becomes $\int_0^x \overline{F}(u)\mathrm{d}u$, that represents the mean time to replacement when the unit is replaced preventively at time x or correctively at failure, whichever takes place first. When $t \to \infty$,

$$\lim_{t\to\infty} \frac{\int_t^{t+x} \overline{F}(u)\mathrm{d}u}{\overline{F}(t)} = \lim_{t\to\infty} \frac{F(t+x) - F(t)}{f(t)} = \lim_{t\to\infty} \frac{\lambda(t; x)}{h(t)} = \frac{1}{h(\infty)}.$$

Differentiating $\int_t^{t+x} \overline{F}(u)\mathrm{d}u / \overline{F}(t)$ with t, and noting that

$$\begin{aligned} &h(t) \int_t^{t+x} \overline{F}(t)\mathrm{d}t - [F(t+x) - F(t)] \\ &\le h(t) \int_t^{t+x} \left[\frac{f(u)}{h(t)}\right] \mathrm{d}u - [F(t+x) - F(t)] = 0. \end{aligned}$$

which shows that $F(t; x)$ decreases with t from $\int_0^x \overline{F}(u)\mathrm{d}u$ to $1/h(\infty)$.

Using $\lambda(t; x)$ and $F(t; x)$, we define:

$$\Lambda(t; x) \equiv \frac{F(t+x) - F(t)}{\int_t^{t+x} \overline{F}(u)\mathrm{d}u} \quad (0 < x < \infty), \tag{4}$$

which means the average failure rate, given that the unit is still survival at time t. It can be easily proved that $\Lambda(t; x)$ increases with t from $F(x)/\int_0^x \overline{F}(u)\mathrm{d}u$ to $h(\infty)$, and $h(t) \le \Lambda(t; x) \le h(t+x)$.

3 Age Replacement

In this section, we apply the above average failure rate into age replacement policies with random arrival of missions. That is, the unit begins to operate after installation, and its failure time X $(0 < X < \infty)$ has a general distribution $F(t) \equiv \Pr\{X \le t\}$ with finite mean $\mu \equiv \int_0^\infty \overline{F}(t)\mathrm{d}t$. In addition, the unit should be active at time T_o $(0 < T_o < \infty)$ for an interval $[T_o, T_o + t_x]$ $(0 \le t_x < \infty)$ to provide reliability. In this case, t_x can be considered as a mission interval during which the unit provides reliability in [2].

3.1 Constant T_o

We plan the unit is replaced preventively at time $T_o + t_x$ $(0 \le t_x \le \infty)$ when it is still survival at time T_o $(0 \le T_o < \infty)$, or it is replaced correctively at failure time X during $(0, T_o + t_x]$, whichever takes place first.

The probability that the unit is replaced at $T_o + t_x$ is

$$\Pr\{X > T_o + t_x\} = \overline{F}(T_o + t_x), \tag{5}$$

and the probability that it is replaced at failure is

$$\Pr\{X \le T_o + t_x\} = F(T_o + t_x). \tag{6}$$

The mean time from installation to replacement is

$$(T_o + t_x)\overline{F}(T_o + t_x) + \int_0^{T_o+t_x} t\,\mathrm{d}F(t) = \int_0^{T_o+t_x} \overline{F}(t)\mathrm{d}t \tag{7}$$

Thus, the expected replacement cost rate is

$$C_s(t_x; T_o) = \frac{c_p + (c_f - c_p)F(T_o + t_x)}{\int_0^{T_o+t_x} \overline{F}(t)\mathrm{d}t}, \tag{8}$$

where c_f and c_p $(c_p < c_f)$ are the costs of replacement policies done at failure and at $T_o + t_x$, respectively.

We find optimum t_x^* to minimize $C_s(t_x; T_o)$ in (8). Differentiating $C_s(t_x; T_o)$ with respect to t_x and setting it equal to zero,

$$h(T_o + t_x) \int_0^{T_o+t_x} \overline{F}(t)\mathrm{d}t - F(T_o + t_x) = \frac{c_p}{c_f - c_p}, \tag{9}$$

whose left-hand side increases with t_x from

$$h(T_o) \int_0^{T_o} \overline{F}(t)\mathrm{d}t - F(T_o)$$

to $h(\infty)/\mu - 1$. Thus, if $h(t)$ increases strictly with t to $h(\infty) = \infty$, then there exists a finite and unique t_x^* $(0 \le t_x^* < \infty)$ which satisfies (9), and the resulting cost rate is

$$C_s(t_x^*; T_o) = (c_f - c_p)h(T_o + t_x^*). \tag{10}$$

Noting that the left-hand side of (9) increases with T_o, t_x^* decreases with T_o from T^* to 0, where T^* is an optimum age replacement time that satisfies

$$h(T)\int_0^T \overline{F}(t)\mathrm{d}t - F(T) = \frac{c_p}{c_f - c_p}. \tag{11}$$

3.2 Random T_o

When T_o is a random variable and has a general distribution $Y(t) \equiv \Pr\{T_o \le t\}$ with a density function $y(t) \equiv \mathrm{d}Y(t)/\mathrm{d}t$ and a finite mean $\gamma = \int_0^\infty \overline{Y}(t)\mathrm{d}t$, we plan that the unit is replaced preventively at time $T_o + t_x$ $(0 \le t_x \le \infty)$ when it is still survival at a random time T_o $(0 \le T_o < \infty)$, or it is replaced correctively at failure time X during $(0, T_o + t_x]$, whichever takes place first.

The probability that the unit is replaced at $T_o + t_x$ is

$$\Pr\{X > T_o + t_x\} = \int_0^\infty \overline{F}(t + t_x)\mathrm{d}Y(t), \tag{12}$$

and the probability that it is replaced at failure is

$$\Pr\{X \le T_o + t_x\} = \int_0^\infty F(t + t_x)\mathrm{d}Y(t). \tag{13}$$

The mean time from installation to replacement is

$$\begin{aligned}&\int_0^\infty (t + t_x)\overline{F}(t + t_x)\mathrm{d}Y(t) + \int_0^\infty \left[\int_0^{t+t_x} u\mathrm{d}F(u)\right]\mathrm{d}Y(t)\\ &= \int_0^\infty \left[\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u\right]\mathrm{d}Y(t).\end{aligned} \tag{14}$$

Thus, the expected replacement cost rate is

$$C_s(t_x; Y) = \frac{c_p + (c_f - c_p)\int_0^\infty F(t + t_x)\mathrm{d}Y(t)}{\int_0^\infty [\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u]\mathrm{d}Y(t)}, \tag{15}$$

where c_f and c_p $(c_p < c_f)$ are the costs of replacement policies done at failure and at $T_o + t_x$, respectively.

Clearly,

$$\lim_{t_x\to\infty} C_s(t_x;Y) = \frac{c_f}{\mu},$$

$$\lim_{t_x\to 0} C_s(t_x;Y) = \frac{c_p + (c_f - c_p)\int_0^\infty F(t)\mathrm{d}Y(t)}{\int_0^\infty \overline{F}(t)\overline{Y}(t)\mathrm{d}t},$$

which agrees with random replacement model in [3].

We find optimum t_x^* to minimize $C_s(t_x;Y)$ in (15). Differentiating $C_s(t_x;Y)$ with respect to t_x and setting it equal to zero,

$$h_s(t_x)\int_0^\infty \left[\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u\right]\mathrm{d}Y(t) - \int_0^\infty F(t+t_x)\mathrm{d}Y(t) = \frac{c_p}{c_f - c_p}, \tag{16}$$

where

$$h_s(t_x) \equiv \frac{\int_0^\infty f(t+t_x)\mathrm{d}Y(t)}{\int_0^\infty \overline{F}(t+t_x)\mathrm{d}Y(t)}.$$

When $Y(t) = 1 - \mathrm{e}^{-\theta t}$,

$$h_s(t_x) \equiv \lim_{T\to\infty} h_f(T;t_x) \equiv \lim_{T\to\infty} \frac{\int_0^T f(t+t_x)\mathrm{d}Y(t)}{\int_0^T \overline{F}(t+t_x)\mathrm{d}Y(t)},$$

and it increases with t_x from $h_s(0) = \int_0^\infty f(t)\mathrm{e}^{-\theta t}\mathrm{d}t / \int_0^\infty \overline{F}(t)\mathrm{e}^{-\theta t}\mathrm{d}t$ to $h(\infty)$. Then, the left-hand side of (16) increases with t_x to ∞ as $h(\infty)\to\infty$. In this case, there exists a finite and unique t_x^* ($0 \le t_x^* < \infty$) which satisfies (16), and the resulting cost rate is

$$C_s(t_x^*;Y) = (c_f - c_p)h_s(t_x^*). \tag{17}$$

When T_o has a gamma distribution with a density function $y(t) = \theta^k t^{k-1}\mathrm{e}^{-\theta t}/(k-1)!$ ($k = 1, 2, \ldots$), and the failure time X has a Weibull distribution $F(t) = 1 - \mathrm{e}^{-(\alpha t)^\beta}$ ($\alpha > 0, \beta > 1$), Table 1 presents optimum t_x^* and its cost rate $C_s(t_x^*;Y)$ for k and c_p when $\theta = 1.0$, $\alpha = 0.1$, $\beta = 2.0$, and $c_f = 100.0$. Table 1 shows that optimum interval $[T_o, T_o + t_x^*]$ decreases with k and increases with c_p. This means that if k becomes large, then the failure rate increases with T_o and t_x^* becomes small. On the other hand, if c_p ($<c_f$) becomes large, then it is unnecessary to replace the unit at a early time and t_o^* becomes large.

Table 1 Optimum t_x^* and its cost rate $C_s(t_x^*; Y)$ when $\theta = 1.0$, $\alpha = 0.1$, $\beta = 2.0$, and $c_f = 100.0$

c_p	$k = 1$		$k = 2$		$k = 5$	
	t_x^*	$C_s(t_x^*; Y)$	t_x^*	$C_s(t_x^*; Y)$	t_x^*	$C_s(t_x^*; Y)$
10.0	2.564	6.269	1.756	6.466	$t_x^* \to 0$	7.048
15.0	3.446	7.397	2.625	7.537	0.147	7.911
20.0	4.282	8.279	3.457	8.385	0.968	8.662
25.0	5.114	8.989	4.286	9.067	1.797	9.276
30.0	5.966	9.565	5.141	9.624	2.659	9.780
35.0	6.863	10.031	6.043	10.076	3.575	10.191
40.0	7.832	10.406	7.018	10.439	4.569	10.521
45.0	8.901	10.700	8.096	10.723	5.671	10.779
50.0	10.111	10.921	9.316	10.937	6.924	10.974

3.3 Replace at T and $T_o + t_x$

In order to prevent early or late arrivals of time T_o, we plan that the unit is replaced preventively at time T $(0 < T \leq \infty)$ or at time $T_o + t_x$ $(0 \leq t_x \leq \infty)$, whichever takes place first. However, no replacement can be done preventively during the interval $[T_o, T_o + t_x]$. In this policy, t_x is constantly given and T_o is a random variable with a general distribution $Y(t)$.

The probability that the unit is replaced at T is

$$\Pr\{X > T, T_o > T\} = \overline{F}(T)\overline{Y}(T), \tag{18}$$

the probability that it is replaced at $T_o + t_x$ is

$$\Pr\{X > T_o + t_x, T_o \leq T\} = \int_0^T \overline{F}(t + t_x)\mathrm{d}Y(t), \tag{19}$$

and the probability that it is replaced at failure is

$$\Pr\{X \leq T \text{ and } T_o \geq T, X \leq T_o + t_x \text{ and } T_o < T\} = F(T)\overline{Y}(T) + \int_0^T F(t + t_x)\mathrm{d}Y(t), \tag{20}$$

where note that (18) + (19) + (20) = 1.

The mean time from installation to replacement is

$$T\overline{F}(T)\overline{Y}(T) + \int_0^T (t+t_x)\overline{F}(t+t_x)\mathrm{d}Y(t) + \overline{Y}(T)\int_0^T t\mathrm{d}F(t)$$
$$+ \int_0^T \left[\int_0^{t+t_x} u\mathrm{d}F(u)\right]\mathrm{d}Y(t)$$
$$= \overline{Y}(T)\int_0^T \overline{F}(t)\mathrm{d}t + \int_0^T \left[\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u\right]\mathrm{d}Y(t). \tag{21}$$

Thus, the expected replacement cost rate is

$$C_f(T;t_x) = \frac{c_p + (c_f - c_p)[F(T)\overline{Y}(T) + \int_0^T F(t+t_x)\mathrm{d}Y(t)]}{\overline{Y}(T)\int_0^T \overline{F}(t)\mathrm{d}t + \int_0^T [\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u]\mathrm{d}Y(t)}, \tag{22}$$

Note that when $t_x \to \infty$, $\lim_{t_x\to\infty} C_f(T;t_x)$ becomes age replacement model in [2], when $t_x \to 0$, $\lim_{t_x\to 0} C_f(T;t_x)$ becomes random replacement model in [3], when $T \to \infty$, $\lim_{T\to\infty} C_f(T;t_x) = C_s(t_x;Y)$ in (15), and when $T \to 0$, $\lim_{T\to 0} C_f(T;t_x) = \infty$.

We find optimum T_f^* and t_{xf}^* to minimize $C_f(T;t_x)$ in (22) for given t_x. Differentiating $C_f(T;t_x)$ with respect to T and setting it equal to zero,

$$q_f(T;t_x)\left\{\overline{Y}(T)\int_0^T \overline{F}(t)\mathrm{d}t + \int_0^T \left[\int_0^{t+t_x} \overline{F}(u)\mathrm{d}u\right]\mathrm{d}Y(t)\right\}$$
$$-\left[F(T)\overline{Y}(T) + \int_0^T F(t+t_x)\mathrm{d}Y(t)\right] = \frac{c_p}{c_f - c_p}, \tag{23}$$

where

$$q_f(T;t_x) \equiv \frac{r(T)\lambda(T;t_x) + h(T)}{r(T)\dfrac{\lambda(T;t_x)}{\Lambda(T;t_x)} + 1} \quad \text{and} \quad r(T) \equiv \frac{y(T)}{\overline{Y}(T)},$$

and the instant failure rate $h(T)$, the conditional failure probability $\lambda(T;t_x)$ and the average failure rate $\Lambda(T;t_x)$ are included in $q_f(T;t_x)$.

When $Y(t) = 1 - \mathrm{e}^{-\theta t}$, $r(T) = \theta$ and

$$q_f(T;t_x) = \frac{\theta[F(T+t_x) - F(T)] + f(T)}{\theta\int_T^{T+t_x} \overline{F}(t)\mathrm{d}t + \overline{F}(T)}.$$

Note that

$$h(T) < \frac{F(T+t_x) - F(T)}{\int_T^{T+t_x} \overline{F}(t)\mathrm{d}t} < h(T+t_x),$$

then $q_f(T;t_x)$ increases strictly with T to ∞ as $h(\infty)\to\infty$, and also increases strictly with t_x to $q_f(T;\infty)$. Thus, the left-hand side of (23) increases with T from 0 to ∞ as $h(\infty)\to\infty$. In this case, there exists a finite and unique T_f^* $(0<T_f^*<\infty)$ which satisfies (23), and the resulting cost rate is

$$C_f(T_f^*;t_x)=(c_f-c_p)q_f(T_f^*;t_x). \tag{24}$$

In addition, the left-hand side of (23) increases with t_x, then T_f^* decreases with t_x from T^* which satisfies the following random replacement model [3],

$$h(T)\int_0^T \mathrm{e}^{-\theta t}\overline{F}(t)\mathrm{d}t-\int_0^T \mathrm{e}^{-\theta t}\mathrm{d}F(t)=\frac{c_p}{c_f-c_p}.$$

Nest, we find optimum t_{xf}^* for given T. Differentiating $C_f(T;t_x)$ with respect to t_x for given T and setting it equal to zero,

$$\begin{aligned}&h_f(T;t_x)\left\{\overline{Y}(T)\int_0^T\overline{F}(t)\mathrm{d}t+\int_0^T\left[\int_0^{t+t_x}\overline{F}(u)\mathrm{d}u\right]\mathrm{d}Y(t)\right\}\\&-\left[F(T)\overline{Y}(T)+\int_0^T F(t+t_x)\mathrm{d}Y(t)\right]=\frac{c_p}{c_f-c_p},\end{aligned} \tag{25}$$

where

$$h_f(T;t_x)\equiv\frac{\int_0^T f(t+t_x)\mathrm{d}Y(t)}{\int_0^T\overline{F}(t+t_x)\mathrm{d}Y(t)}<h(T+t_x).$$

When $Y(t)=1-\mathrm{e}^{-\theta t}$, $h_f(T;t_x)$ increases with t_x to $h(\infty)$. Then, the left-hand side of (25) increases strictly with t_x from 0 to ∞ as $h(\infty)\to\infty$. In this case, there exists a finite and unique t_{xf}^* $(0<t_{xf}^*<\infty)$ which satisfies (25), and the resulting cost rate is

$$C_f(T;t_{xf}^*)=(c_f-c_p)h_f(T;t_{xf}^*). \tag{26}$$

Note that t_{xf}^* decreases with T to t_x^* given in (16), as the left-hand side of (22) increases with T to that of (16).

When $y(t)=\theta^k t^{k-1}\mathrm{e}^{-\theta t}/(k-1)!$ $(k=1,2,\ldots)$ and $F(t)=1-\mathrm{e}^{-(\alpha t)^\beta}$, $(\alpha>0,\beta\ge1)$, Table 2 presents optimum T_f^* and its cost rate $C_f(T_f^*;t_x)$ for t_x and c_p when $\theta=1.0, k=2, \alpha=0.1, \beta=2.0$, and $c_f=100.0$, and Table 3 presents optimum t_{xf}^* and its cost rate $C_f(T;t_{xf}^*)$ for T and c_p when $\theta=1.0, k=2, \alpha=0.1, \beta=2.0$, and $c_f=100.0$. Table 3 shows that T_f^* increases with c_p and decreases with t_x and t_{xf}^* increases with c_p and decreases with T, as shown in the above analytical discussions.

Table 2 Optimum T_f^* and its cost rate $C_f(T_f^*; t_x)$ when $\theta = 1.0$, $k = 2$, $\alpha = 0.1$, $\beta = 2.0$, and $c_f = 100.0$

c_p	$t_x = 1.0$		$t_x = 2.0$		$t_x = 5.0$	
	T_f^*	$C_f(T_f^*; t_x)$	T_f^*	$C_f(T_f^*; t_x)$	T_f^*	$C_f(T_f^*; t_x)$
10.0	3.368	6.442	2.864	6.174	2.167	6.977
15.0	4.577	8.148	3.834	7.493	2.871	7.786
20.0	5.888	9.769	4.864	8.702	3.587	8.453
25.0	7.354	11.360	6.001	9.861	4.357	9.048
30.0	9.027	12.943	7.291	11.002	5.212	9.602
35.0	10.961	14.524	8.776	12.136	6.185	10.135
40.0	13.260	16.105	10.513	13.268	7.315	10.659
45.0	16.050	17.686	12.612	14.400	8.650	11.179
50.0	19.265	19.266	15.380	15.532	10.257	11.698

Table 3 Optimum t_{xf}^* and its cost rate $C_f(T; t_{xf}^*)$ when $\theta = 1.0$, $k = 2$, $\alpha = 0.1$, $\beta = 2.0$, and $c_f = 100.0$

c_p	$T = 1.0$		$T = 2.0$		$T = 5.0$	
	t_{xf}^*	$C_f(T; t_{xf}^*)$	t_{xf}^*	$C_f(T; t_{xf}^*)$	t_{xf}^*	$C_f(T; t_{xf}^*)$
10.0	3.918	8.137	2.446	6.330	1.800	6.350
15.0	5.541	10.440	3.512	7.781	2.673	7.447
20.0	7.135	12.374	4.543	8.964	3.510	8.317
25.0	8.776	14.058	5.581	9.953	4.345	9.022
30.0	10.523	15.564	6.659	10.791	5.206	9.598
35.0	12.442	16.943	7.811	11.510	6.116	10.070
40.0	14.606	18.234	9.069	12.126	7.101	10.452
45.0	17.113	19.468	10.479	12.659	8.190	10.754
50.0	20.088	20.670	12.101	13.122	9.424	10.985

4 Minimal Repair

It is assumed that the unit undergoes minimal repairs at failures and begins to operate again after repairs, where the time for repairs are negligible and the failure rate remains undisturbed by repairs. In this case, we define

$$\Lambda(t; x) = \frac{1}{x} \int_t^{t+x} h(u)\mathrm{d}u, \tag{27}$$

which means the average failure rate for an interval $[t, t + x]$. It is obviously to show that $\Lambda(t; x)$ increases with t from $H(x)/x$ to $h(\infty)$ and increases with x from $h(0)$ to $h(\infty)$, and $h(t) \leq \Lambda(t; x) \leq h(t + x)$.

4.1 Constant T_o

In order to prevent an increasing repair cost, we plan that the unit is replaced at time $T_o + t_x$ $(0 < T_o \leq \infty, 0 \leq t_x < \infty)$. Noting that the expected number of failures during $(0, T_o + t_x]$ is $H(T_o + t_x)$, the expected cost rate is

$$C_s(t_x; T_o) = \frac{c_m H(T_o + t_x) + c_p}{T_o + t_x}, \tag{28}$$

where c_m is minimal repair cost at failure, and c_p is given in (8).

We find optimum t_x^* to minimize $C_s(t_x; T_o)$ for given T_o. Differentiating $C_s(t_x; T_o)$ with respect to t_x and setting it equal to zero,

$$h(T_o + t_x)(T_o + t_x) - H(T_o + t_x) = \frac{c_p}{c_m}, \tag{29}$$

whose left-hand side increases with t_x from $h(T_o)T_o - H(T_o)$ to $\int_0^\infty [h(\infty) - h(t)]\mathrm{d}t$. Thus, if the failure rate $h(t)$ increases strictly with t to $h(\infty) = \infty$, then there exists a finite and unique t_x^* $(0 \leq t_x^* < \infty)$ which satisfies (29), and the resulting cost rate is

$$C_s(t_x^*; T_o) = c_m h(T_o + t_x^*), \tag{30}$$

Noting that the left-hand side of (29) increases with T_o, t_x^* decreases with T_o from T^* to 0, where T^* is an optimum periodic replacement time that satisfies

$$h(T)T - H(T) = \frac{c_p}{c_m}. \tag{31}$$

4.2 Random T_o

We plan that the unit is replaced at time $T_o + t_x$ $(0 \leq t_x < \infty)$, where T_o is a random variable with distribution $Y(t)$. Then, the expected cost rate is

$$C_s(t_x; Y) = \frac{c_m \int_0^\infty H(t + t_x)\mathrm{d}Y(t) + c_p}{\int_0^\infty (t + t_x)\mathrm{d}Y(t)}, \tag{32}$$

where c_m is minimal repair cost at failure, and c_p is given in (15).

Clearly, $\lim_{t_x \to \infty} C_s(t_x; Y) \to \infty$ and

$$\lim_{t_x \to 0} C_s(t_x; Y) = \frac{c_m \int_0^\infty H(t)\mathrm{d}Y(t) + c_p}{\int_0^\infty t\mathrm{d}Y(t)},$$

which agrees with random replacement model [3].

If there exists an optimum t_x^* to minimize $C_s(t_x; Y)$ in (32), it satisfies

$$\int_0^\infty (t+t_x)\mathrm{d}Y(t) \int_0^\infty h(t+t_x)\mathrm{d}Y(t) - \int_0^\infty H(t+t_x)\mathrm{d}Y(t) = \frac{c_p}{c_m}, \tag{33}$$

whose left-hand side increases with t_x to ∞ as $h(\infty) \to \infty$. In this case, the resulting cost rate is

$$C_s(t_x^*; Y) = c_m \int_0^\infty h(t+t_x^*)\mathrm{d}Y(t). \tag{34}$$

When $y(t) = \theta^k t^{k-1} \mathrm{e}^{-\theta t}/\Gamma(k)$ and $F(t) = 1 - \mathrm{e}^{-(\alpha t)^\beta}$, Table 4 presents optimum t_x^* and its cost rate $C_s(t_x^*; Y)$ for k and c_m when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, and $c_p = 100.0$. Table 4 shows that optimum interval $[T_o, T_o + t_x^*]$ decreases when c_m increases and T_o arrives at a late time due to the total increasing repair cost. Note that when $k = 5$, $t_x^* \to 0$ for all of c_m.

4.3 Replace at T and $T_o + t_x$

We plan that the unit is replaced at time T $(0 < T \le \infty)$ or at time $T_o + t_x$ $(0 \le t_x \le \infty)$, whichever takes place first; however, only minimal repairs can be done during the interval $[T_o, T_o + t_x]$. Then, the expected number of repairs between replacement policies is

$$H(T)\overline{Y}(T) + \int_0^T H(t+t_x)\mathrm{d}Y(t), \tag{35}$$

Table 4 Optimum t_x^* and its cost rate $C_s(t_x^*; Y)$ when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, and $c_p = 100.0$

c_m	$k = 1$		$k = 2$		$k = 5$	
	t_x^*	$C_s(t_x^*; Y)$	t_x^*	$C_s(t_x^*; Y)$	t_x^*	$C_s(t_x^*; Y)$
10.0	2.317	66.324	1.465	69.178	$t_x^* \to 0$	77.383
15.0	1.769	83.324	0.944	88.133	$t_x^* \to 0$	105.355
20.0	1.449	97.953	0.644	105.529	$t_x^* \to 0$	133.327
25.0	1.236	111.782	0.447	122.048	$t_x^* \to 0$	161.300
30.0	1.081	124.855	0.307	138.057	$t_x^* \to 0$	189.272
35.0	0.963	137.392	0.201	153.651	$t_x^* \to 0$	217.244
40.0	0.871	149.613	0.118	168.977	$t_x^* \to 0$	245.217
45.0	0.794	161.407	0.051	184.072	$t_x^* \to 0$	273.189
50.0	0.732	173.128	$t_x^* \to 0$	199.001	$t_x^* \to 0$	301.161

and the mean time from installation to replacement is

$$T\overline{Y}(T) + \int_0^T (t + t_x)\mathrm{d}Y(t) = t_x Y(T) + \int_0^T \overline{Y}(t)\mathrm{d}t. \tag{36}$$

Thus, the expected replacement cost rate is

$$C_f(T; t_x) = \frac{c_m[H(T)\overline{Y}(T) + \int_0^T H(t + t_x)\mathrm{d}Y(t)] + c_p}{t_x Y(T) + \int_0^T \overline{Y}(t)\mathrm{d}t}. \tag{37}$$

Differentiating $C_f(T; t_x)$ with respect to T and setting it equal to zero,

$$q_f(T; t_x)\left[t_x Y(T) + \int_0^T \overline{Y}(t)\mathrm{d}t\right] - \left[H(T)\overline{Y}(T) + \int_0^T H(t + t_x)\mathrm{d}Y(t)\right] = \frac{c_p}{c_m}, \tag{38}$$

where

$$q_f(T; t_x) \equiv \frac{r(T)\Lambda(T; t_x) + h(T)/t_x}{r(T) + 1/t_x}. \tag{39}$$

When $Y(t) = 1 - \mathrm{e}^{-\theta t}$, $q_f(T; t_x)$ increases with T to $h(\infty)/(\theta t_x + 1)$. Then, the left-hand side of (38) increases with T from 0 to ∞ as $h(\infty) \to \infty$. Therefore, there exits a finite and unique T_f^* $(0 < T_f^* < \infty)$ which satisfies (38), and the resulting cost rate is

$$C_f(T_f^*; t_x) = c_m \frac{\theta\Lambda(T_f^*; t_x) + h(T_f^*)/t_x}{\theta + 1/t_x}. \tag{40}$$

Next, differentiating $C_f(T; t_x)$ with respect to t_x and setting it equal to zero,

$$\frac{\int_0^T h(t + t_x)\mathrm{d}Y(t)}{Y(T)}\left[t_x Y(T) + \int_0^T \overline{Y}(t)\mathrm{d}t\right] - \left[H(T)\overline{Y}(T) + \int_0^T H(t + t_x)\mathrm{d}Y(t)\right] = \frac{c_p}{c_m}, \tag{41}$$

whose left-hand side increases with t_x to ∞ as $h(\infty) \to \infty$. Therefore, there exists a finite and unique t_{xf}^* $(0 \le t_{xf}^* < \infty)$ which satisfies (40), and the resulting cost rate is

$$C_f(T; t_{xf}^*) = c_m \frac{\int_0^T h(t + t_{xf}^*)\mathrm{d}Y(t)}{Y(T)}. \tag{42}$$

Table 5 Optimum T_f^* and its cost rate $C_f(T_f^*; t_x)$ when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, $t_x = 1.0$, and $c_p = 100.0$

c_m	$k = 1$		$k = 2$		$k = 3$	
	T_f^*	$C_f(T_f^*; t_x)$	T_f^*	$C_f(T_f^*; t_x)$	T_f^*	$C_f(T_f^*; t_x)$
10.0	3.467	74.336	3.115	66.613	3.057	64.619
15.0	2.588	85.137	2.451	79.764	2.451	78.299
20.0	2.139	95.547	2.080	91.265	2.119	90.667
25.0	1.846	104.785	1.846	102.121	1.885	101.321
30.0	1.631	112.852	1.670	111.739	1.709	110.321
35.0	1.494	122.090	1.533	120.521	1.592	120.077
40.0	1.377	130.156	1.416	128.062	1.475	127.323
45.0	1.279	137.637	1.338	136.789	1.396	135.778
50.0	1.201	145.117	1.260	143.873	1.318	142.548

Table 6 Optimum t_{xf}^* and its cost rate $C_f(T; t_{xf}^*)$ when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, $T = 1.0$ and $c_p = 100.0$

c_m	$k = 1$		$k = 2$		$k = 3$	
	t_{xf}^*	$C_f(T; t_{xf}^*)$	t_{xf}^*	$C_f(T; t_{xf}^*)$	t_{xf}^*	$C_f(T; t_{xf}^*)$
10.0	3.096	70.275	3.564	83.445	4.189	97.977
15.0	2.393	84.318	2.588	95.870	2.861	107.121
20.0	1.982	96.018	2.041	105.952	2.158	114.704
25.0	1.709	106.350	1.689	114.862	1.709	120.919
30.0	1.514	115.902	1.436	122.600	1.416	127.524
35.0	1.357	124.281	1.240	129.362	1.182	132.372
40.0	1.221	131.098	1.084	135.342	1.025	138.782
45.0	1.123	138.696	0.967	141.713	0.889	143.825
50.0	1.045	146.294	0.869	147.693	0.791	150.040

When $y(t) = \theta^k t^{k-1} \mathrm{e}^{-\theta t} / \Gamma(k)$ and $F(t) = 1 - \mathrm{e}^{-(\alpha t)^\beta}$, Table 5 presents optimum T_f^* and its cost rate $C_f(T_f^*; t_x)$ for t_x and c_m when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, $t_x = 1.0$, and $c_p = 100.0$, and Table 6 presents optimum t_{xf}^* and its cost rate $C_f(T; t_{xf}^*)$ for k and c_m when $\theta = 1.0$, $\alpha = 1.0$, $\beta = 2.0$, $T = 1.0$, and $c_p = 100.0$.

5 Conclusions

We have firstly obtained a definition of average failure rate, i.e., $\Lambda(t; x)$, that is based on the conditional failure probability and the mean time to failure given that the unit is still survival at time t. The mathematical monotonicity of $\Lambda(t; x)$ has been proved analytically. Next, the average failure rate has been applied into preventive replace-

ment policies when the arrival time of a mission is a random variable and lasts for an interval, during which, the unit provides reliability and no maintenance can be done. Optimum replacement time and mission interval have been discussed respectively for the models of age replacement and periodic replacement. Numerical examples have been illustrated when the mission arrival time follows a gamma distribution and the failure time of the unit has a Weibull distribution.

Acknowledgements This work is supported by National Natural Science Foundation of China (NO. 71801126), Natural Science Foundation of Jiangsu Province (NO. BK20180412), Aeronautical Science Foundation of China (NO. 2018ZG52080), Fundamental Research Funds for the Central Universities (NO. NR2018003), and Japan Society for the Promotion of Science KAKENHI (NO. 18K01713).

References

1. Barlow RE, Proschan F (1965) Mathematical theory of reliability. Wiley
2. Nakagawa T (2005) Mathematical theory of reliability. Springer
3. Nakagawa T (2014) Random maintenance policies. Springer
4. Park M, Pham H (2016) Cost models for age replacement policies and block replacement policies under warranty. Appl Math Model 40:5689–5702
5. Zhao X, Nakagawa T (2018) Advanced maintenance policies for shock and damage models. Springer
6. Sheu SH, Liu TH, Zhang ZG, Tsai HN (2018) The generalized age maintenance policies with random working times. Reliab Eng Syst Saf 169:503–514
7. Chang CC, Chen YL (2019) Optimization of continuous and discrete scheduled times for a cumulative damage system with age-dependent maintenance. Commun Stat Theory Methods 48:4261–4277
8. Nakagawa T, Zhao X (2015) Maintenance overtime policies in reliability theory. Springer
9. Chen M, Zhao X, Nakagawa T (2019) Replacement policies with general models. Ann Oper Res 277:47–61
10. Zhao X, Nakagawa T (2012) Optimization problems of replacement first or last in reliability theory. Eur J Oper Res 223:141–149
11. Hamidi M, Szidarovszky F, Szidarovszky M (2016) New one cycle criteria for optimizing preventive replacement policies. Reliab Eng Syst Saf 154:42–48
12. Okamura H, Dohi T (2017) Moment-based approach for some age-based replacement problems. J Ind Prod Eng 34:558–567
13. Zhao X, Al-Khalifa KN, Hamouda AMS, Nakagawa T (2017) Age replacement models: a summary with new perspectives and methods. Reliab Eng Syst Saf 161:95–105
14. Mannai N, Gasmi S (2018) Optimal design of k-out-of-n system under first and last replacement in reliability theory. Oper Res. https://doi.org/10.1007/s12351-018-0375-4
15. Sheu SH, Liu TH, Zhang ZG (2019) Extended optimal preventive replacement policies with random working cycle. Reliab Eng Syst Saf 188:398–415

Xufeng Zhao is a Professor at Nanjing University of Aeronautics and Astronautics, China. He received his bachelor's degree in information management and information system in 2006, and master's degree in system engineering in 2009, both from Nanjing Tech University, China; and his doctoral degree in business administration and computer science in 2013 from Aichi Institute of Technology, Japan. Dr. Zhao has worked as Postdoctoral Researcher from 2013 to 2017 at Aichi Institute of Technology and Qatar University, respectively. Dr. Zhao is interested in probability

theory, stochastic process, reliability and maintenance theory, and applications in computer and industrial systems. He has published two books in maintenance theory from Springer and more than fifty research papers in peer reviewed journals; and he is the author or coauthor of twelve book chapters from Springer and Wiley, and etc. He has gotten one best paper award from IEEE Reliability Society Japan Chapter and five best paper awards from International conferences in reliability, maintainability and Quality.

Jiajia Cai is a Ph.D. candidate in quality and reliability management at Nanjing University of Aeronautics and Astronautics, China. She received her bachelor's degree in industrial engineering in 2017, and master's degree in management science and engineering in 2020, both from Nanjing University of Aeronautics and Astronautics, China. Jiajia Cai is interested in reliability theory and its applications in industrial systems.

Satoshi Mizutani received Ph.D. degree from Aichi Institute of Technology in 2004, Japan. He was a visiting researcher at Kinjo Gakuin University in Nagoya City, from 2004 to 2007, Japan. He worked as Assistant Professor from 2007 to 2013, and as Associate Professor from 2013 to 2018 at Aichi University of Technology, Japan. He is now Associate Professor at Aichi Institute of Technology, Japan. His research interests are optimization problems of inspection policies for computer systems in reliability theory. Dr. Mizutani has received Outstanding Young Scientist Award, best paper awards from IEEE Reliability Society Japan Chapter and the 18th Asia Pacific Conference on Industrial Engineering and Management Systems.

Toshio Nakagawa received B.S.E. and M.S. degrees from Nagoya Institute of Technology in 1965 and 1967, Japan, respectively; and a Ph.D. degree from Kyoto University in 1977, Japan. He worked as a Research Associate at Syracuse University for two years from 1972 to 1973. He is now a Honorary Professor with Aichi Institute of Technology, Japan. He has published 5 books from Springer, and about 200 journal papers. His research interests are optimization problems in operations research and management science, and analysis for stochastic and computer systems in reliability and maintenance theory.

Optimal Maintenance Models of Social Infrastructures Considering Natural Disasters

Takumi Kishida, Kodo Ito, Higuchi Yoshiyuki and Toshio Nakagawa

Abstract Social infrastructures such as roads and bridges in Japan are to exceed 50 years after their constructions in 2023. These infrastructures have to be maintained under strict budgets of local governments because local governments maintain them and are facing serious financial difficulties by the downturn of local economies. Moreover, they have to be maintained under scarcity of skillful engineers because of the aging workforce. Comparing mechanical system maintenance, social infrastructure maintenances have some practical interesting characteristics such as maintenance time delay and wide variety of preventive maintenance costs. Maintenance policies with various kinds of factors such as maintenance periods, maintenance time delay, wide variety of preventive maintenance costs and degradation levels, and natural disaster distribution, have to be practically established. In this chapter, considering natural disasters, we form stochastically cumulative damage models and discuss their optimal policies theoretically and numerically. The expected cost rates are obtained and optimal preventive maintenance levels which minimize them are derived. Extended models with a natural disaster and its recovery based on these models could be proposed and be applied to actual social infrastructures.

T. Kishida (✉) · K. Ito
Department of Social Management Engineering, Graduate School of Engineering, Tottori University, 4-101 Koyama-cho Minami, Tottori 680-8550, Japan
e-mail: M19J6009A@edu.tottori-u.ac.jp

K. Ito
e-mail: itokodo@tottori-u.ac.jp

H. Yoshiyuki
Faculty of Symbiotic Systems Science, Fukushima University, Kanayagawa, Fukushima 960-1248, Japan
e-mail: yosh@higutic.org

T. Nakagawa
Department of Business Administration, Aichi Institute of Technology, 1247 Yachigusa, Yakusa Cho, Toyota City, Aichi 470-0392, Japan
e-mail: nakagawa@aitech.ac.jp

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_15

Keywords Social infrastructure · Damage model · Maintenance policy · Natural disaster · Preventive maintenance

1 Introduction

In Japan, the ages of 39% of road bridges and 27% of tunnels are to reach 50 years of construction in 2023, and large-scale maintenances of them have become a big national issue of Japanese government because such social infrastructures are fundamental and indispensable for our lives and economic activities [1]. Annual inspections and maintenances of roads are performed by local governments which have serious financial difficulties because of the lower fertility rate, the aging workforce, and the downturn of local economy. So, efficient maintenance methods and plans are always necessary.

Because infrastructures deteriorate with time stochastically and maintenance of them are big issues, lots of studies concerning degradation prediction methods of roads and their optimal maintenance policies have been performed. In case of the highway rehabilitation, Friesz et al. and Markow et al. considered optimal rehabilitation frequencies of a single highway pavement [2, 3]. Tsunokawa et al. considered optimal rehabilitation problem for multiple rehabilitations using the discontinuous pavement condition which is approximated by a linear deterioration rate continuous function [4]. Li et al. extended this consideration using Markov decision process (MDP) [5]. Ouyang et al. considered optimal rehabilitation scheduling of multiple highway pavement as a mixed-integer nonlinear program and derived optimal rehabilitation timing for a single pavement analytically [6, 7]. Kobayshi et al. expressed the deterioration process of the pavement by the combined MDP model of the deterioration process of the road surface integrity and the deterioration process of the load capacity of the pavement, and proposed the optimal repair and update model to minimize the expected Life Cycle Cost [8]. In case of bridge and tunnel maintenance, Kaito et al. pointed out that LCC can be separated average costs and relative costs in the case of bridge maintenance [9]. Sakai et al. proposed the model to predict deterioration of tunnel structures using Poisson process [10]. Obama et al. proposed a methodology for designing optimal scrapping and maintenance policies for infrastructures that are to be demolished due to the demand decline or the deterioration progression [11]. Tsuda et al. proposed a methodology to estimate the Markovian transition probability model to forecast the deterioration process of bridge components [12]. Authors considered several optimal maintenance policies for infrastructures using cumulative damage model and Markov chain model [13–15].

Because the annual budget of local government has a limitation, some preventive maintenance works might be delayed sometimes. Although maintenance costs are assumed to be constant, they change with damage levels in practice. We consider optimal maintenance models of social infrastructures applying cumulative damage model. We form the delayed maintenance model and introduce modified cost ones. In case of modified cost models, using modified and variable costs, expected cost

rates are obtained, and optimal damage levels which minimize them are discussed. Basic and expanded models with disaster recovery are discussed.

2 Basic Models

First, we denote the fundamental model as follows [16, p. 45]:

(1) Social infrastructures such as roads and bridges are degraded with time by traffics, and its amount of damage is measured only at time jT $(j = 1, 2, \ldots)$ for a specified $T > 0$. Each amount W_j of damage between periodic times has an identical distribution $G(x) \equiv \mathrm{P_r}\{W_j \leq x\}$ with mean time $1/\mu$.
(2) When the total damage exceeds a threshold level Z_K at time jT, the corrective maintenance(CM) is done at time jT, and when it is between Z_1 $(Z_1 < Z_K)$ and Z_K, the preventive maintenance(PM) is done at time jT.
(3) CM cost is c_K and PM cost is $c_1 (< c_K)$.

The expected cost rate is

$$TC_0(Z_1) = \frac{c_K - (c_K - c_1)\left[G(Z_K) - \int_0^{Z_1} \overline{G}(Z_K - x)\mathrm{d}M_G(x)\right]}{1 + M_G(Z_1)}, \tag{1}$$

where $M_G(x) \equiv \sum_{j=1}^{\infty} G^{(j)}(x)$, $\phi^{(j)}(x)$ $(j = 1, 2, \ldots)$ denotes the j-fold Stieltjes convolution of distribution $\phi(x)$, and $\phi^{(0)}(x) \equiv 1$ for $x \geq 0$. Differentiating $C_0(Z_1)$ with respect to Z_1 and putting it equal to zero,

$$\int_{Z_K - Z_1}^{Z_K} [1 + M_G(Z_K - x)]\,\mathrm{d}G(x) = \frac{c_1}{c_K - c_1}, \tag{2}$$

whose left-hand side is strictly increasing from 0 to $M_G(Z_K)$. Thus, if $M_G(Z_K) > c_1/(c_K - c_1)$, then there exists a finite and unique Z_1^* $(0 < Z_1^* < Z_K)$ that satisfies (2), and its resulting cost rate is

$$TC_0(Z_1^*) = (c_K - c_1)\overline{G}(Z_K - Z_1^*). \tag{3}$$

Conversely, if $M_G(Z_K) \leq c_1/(c_K - c_1)$, then $Z_1^* = Z_K$, i.e., the unit should be replaced only at failure, and expected cost rate is

$$TC_0(Z_K) = \frac{c_K}{1 + M_G(Z_K)}. \tag{4}$$

We proposed three models which are expanded ones of the above fundamental model. The delayed maintenance is considered in Model 1. Multiple PM costs

are introduced in Model 2, and multiple PM and additional replacement costs are introduced in Model 3.

2.1 Model 1

We make assumptions (1) and (3) and rewrite (2) as follows:

(2') When the total damage exceeds a threshold level Z_K at time jT, CM is done at time jT, and when it is between Z_1 and Z_K, PM is done at time $(j+1)T$.

The probability that the total damage is between Z_1 and Z_K at time $(j+1)T$ and at time $(j+2)T$, given that it is below Z_1 at time jT, is

$$\begin{aligned}
&\sum_{j=0}^{\infty}\int_0^{Z_1}\left[\int_0^{Z_K-x} G(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\\
&-\sum_{j=0}^{\infty}\int_0^{Z_1}\left[\int_0^{Z_1-x} G(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\\
&=\sum_{j=0}^{\infty}\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x} G(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x), \qquad (5)
\end{aligned}$$

and the probability that it is between Z_1 and Z_K at time $(j+1)T$, given that it is below Z_1 at time jT, and exceeds Z_K at time $(j+2)T$ is

$$\sum_{j=0}^{\infty}\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x} \overline{G}(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\,. \qquad (6)$$

Thus, the mean time to maintenance is

$$\begin{aligned}
&\sum_{j=0}^{\infty}(j+1)T\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}(j+2)T\int_0^{Z_1}\left[G(Z_K-x)-G(Z_1-x)\right]\mathrm{d}G^{(j)}(x)\\
&=T\left[1+G(Z_K)+\int_0^{Z_1}G(Z_K-x)\mathrm{d}M_G(x)\right], \qquad (7)
\end{aligned}$$

and the total expected cost until maintenance is

$$
\begin{aligned}
&c_K \sum_{j=0}^{\infty} \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x) \\
&+c_1 \sum_{j=0}^{\infty} \int_0^{Z_1} \left[\int_{Z_1 - x}^{Z_K - x} G(Z_K - x - y) \mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x) \\
&+c_K \sum_{j=0}^{\infty} \int_0^{Z_1} \left[\int_{Z_1 - x}^{Z_K - x} \overline{G}(Z_K - x - y) \mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x) \\
&= c_K - (c_K - c_1) \left\{ \int_{Z_1}^{Z_K} G(Z_K - y) \mathrm{d}G(y) \right. \\
&\left. + \int_0^{Z_1} \left[\int_{Z_1 - x}^{Z_K - x} G(Z_K - x - y) \mathrm{d}G(y) \right] \mathrm{d}M_G(x) \right\}. \qquad (8)
\end{aligned}
$$

Thus, the expected cost rate is, from (7) and (8) [16, p. 57],

$$
TC_1(Z_1) = \frac{c_K - (c_K - c_1) \left\{ \int_{Z_1}^{Z_K} G(Z_K - y) \mathrm{d}G(y) + \int_0^{Z_1} \left[\int_{Z_1 - x}^{Z_K - x} G(Z_K - x - y) \mathrm{d}G(y) \right] \mathrm{d}M_G(x) \right\}}{1 + G(Z_K) + \int_0^{Z_1} G(Z_K - x) \mathrm{d}M_G(x)}. \qquad (9)
$$

Assuming $G(x) = 1 - \exp(-\mu x)$, (9) is rewritten as,

$$
TC_1(Z_1) = \frac{c_K - (c_K - c_1) \left\{ 1 - [1 + \mu(Z_K - Z_1)] \mathrm{e}^{-\mu(Z_K - Z_1)} \right\}}{1 + \mu Z_1 + 1 - \mathrm{e}^{-\mu(Z_K - Z_1)}}. \qquad (10)
$$

We find optimal Z_1^* to minimize $C_1(Z_1^*)$. Differentiating $C_1(Z_1)$ with respect to Z_1 and putting it equal to zero,

$$
\mathrm{e}^{-\mu(Z_K - Z_1)} \left[\frac{\mu(1 + \mu Z_1)(Z_K - Z_1)}{1 - \mathrm{e}^{-\mu(Z_K - Z_1)}} - 1 \right] = \frac{c_1}{c_K - c_1}. \qquad (11)
$$

Letting $Q_1(Z_1)$ denote the left-hand side of (11), $Q_1(Z_1)$ is strictly increasing in Z_1, and

$$
Q_1(0) = \frac{\mu Z_K - \left(1 - \mathrm{e}^{-\mu Z_K}\right)}{\mathrm{e}^{\mu Z_K} - 1}, \quad Q_1(Z_K) = \mu Z_K.
$$

Thus, if

$$
\frac{\mu Z_K - \left(1 - \mathrm{e}^{-\mu Z_K}\right)}{\mathrm{e}^{\mu Z_K} - 1} < \frac{c_1}{c_K - c_1} < \mu Z_K,
$$

then there exists a finite and unique Z_1^* $(0 < Z_1^* < Z_K)$ which satisfies (11).

2.2 Model 2

We make assumption (1) of Model 1 and rewrite assumptions (2) and (3) as follows:

(2") Degradation states are denoted as discrete degradation levels Z_i ($i = 1, 2, \ldots, K-1, K$), where $Z_1 < Z_2 < \cdots < Z_K$ for $K \geq 2$. When the total damage exceeds a threshold level Z_K at time jT, CM is done at time jT, and when it is between Z_1 and Z_K, PM is done at time jT.

(3') CM cost is c_K and PM cost is c_i ($< c_{i+1}$) when the total damage is between Z_i and Z_{i+1} ($i = 1, 2, \ldots, K-1$).

The probability that CM is done at time $(j+1)T$ is

$$\mathrm{P}_K = \sum_{j=0}^{\infty} \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x)\,, \tag{12}$$

and the probabilities that PM is done at time $(j+1)T$ are

$$\mathrm{P}_i = \sum_{j=0}^{\infty} \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}G^{(j)}(x) \quad (i = 1, 2, \ldots, K-1). \tag{13}$$

The mean time to maintenance is

$$\begin{aligned} &\sum_{j=0}^{\infty} (j+1)T \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x) \\ &+ \sum_{j=0}^{\infty} (j+1)T \sum_{i=1}^{K-1} \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}G^{(j)}(x) = T\left[1 + M_G(Z_1)\right]. \end{aligned} \tag{14}$$

The total expected cost until maintenance is

$$\begin{aligned} \sum_{i=1}^{K} c_i \mathrm{P}_i = c_K - \sum_{i=1}^{K-1} (c_K - c_i) \Bigg\{ & G(Z_{i+1}) - G(Z_i) \\ & + \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}M_G(x) \Bigg\}. \end{aligned} \tag{15}$$

Thus, the expected cost rate is, from (14) and (15) [16, p. 59],

$$TC_2(Z_1) = \frac{c_K - \sum_{i=1}^{K-1} (c_K - c_i) \left\{ G(Z_{i+1}) - G(Z_i) + \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}M_G(x) \right\}}{1 + M_G(Z_1)}. \tag{16}$$

Because PM is done when degradation level is greater than Z_1, $TC_2(Z_1)$ is expressed as a function of Z_1. We find optimal Z_1^* to minimize $C_2(Z_1)$. Differentiating $C_2(Z_1)$ with respect to Z_1 and putting it equal to zero,

$$\sum_{i=2}^{K}(c_i - c_{i-1}) \int_{Z_i - Z_1}^{Z_i} [1 + M_G(Z_i - x)]\,\mathrm{d}G(x) = c_1\,. \tag{17}$$

Letting denote the left-hand side of Eq. (17) by $Q_2(Z_1)$,

$$Q_2(0) = 0\,,$$
$$Q_2(Z_2) = (c_2 - c_1)M_G(Z_2) + \sum_{i=3}^{K}(c_i - c_{i-1}) \int_{Z_i - Z_2}^{Z_i} [1 + M_G(Z_i - x)]\mathrm{d}G(x)\,.$$

Thus, if $Q_2(Z_2) > c_1$, then there exists a finite and unique Z_1^* $(0 < Z_1^* < Z_2)$ which satisfies (17). In particular, when $K = 2$, if $M_G(Z_2) > c_1/(c_2 - c_1)$, then a finite Z_1^* exists.

3 Model 3

We make assumptions (1) of Model 1 and (2") of Model 2 and rewrite assumption (3) as follows:

(3") CM cost is $c_K + c_0(Z_K)$ and PM cost is $c_i + c_0(x)$ when the total damage is between Z_i and Z_{i+1} $(i = 1, 2, \ldots, K-1)$, where $c_0(x)$ is an additional replacement cost that is variable for the total damage x with $c_0(0) = 0$.

The expected cost until maintenance is

$$\begin{aligned}
&[c_K + c_0(Z_K)] \sum_{j=0}^{\infty} \int_0^{Z_1} \overline{G}(Z_K - x)\mathrm{d}G^{(j)}(x) \\
&+ \sum_{i=1}^{K-1} \sum_{j=0}^{\infty} \int_0^{Z_1} \left\{ \int_{Z_i - x}^{Z_{i+1} - x} [c_i + c_0(x + y)]\mathrm{d}G(y) \right\} \mathrm{d}G^{(j)}(x) \\
&= c_K - \sum_{i=1}^{K-1} (c_K - c_i) \left\{ G(Z_{i+1}) - G(Z_i) + \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x) \right] \mathrm{d}M_G(x) \right\} \\
&+ \sum_{j=0}^{\infty} \int_0^{Z_1} \left[\int_x^{Z_K} \overline{G}(y - x)\mathrm{d}c_0(y) \right] \mathrm{d}G^{(j)}(x)\,.
\end{aligned} \tag{18}$$

Thus, the expected cost rate is, from (14) and (18) [16, p. 48],

$$TC_3(Z_1) = \frac{\begin{array}{l} c_K - \sum_{i=1}^{K-1}(c_K - c_i)\Big\{G(Z_{i+1}) - G(Z_i) \\ \quad + \int_0^{Z_1}\left[G(Z_{i+1}-x) - G(Z_i - x)\right]\mathrm{d}M_G(x)\Big\} \\ \quad + \sum_{j=0}^{\infty}\int_0^{Z_1}\left[\int_x^{Z_K}\overline{G}(y-x)\mathrm{d}c_0(y)\right]\mathrm{d}G^{(j)}(x) \end{array}}{1 + M_G(Z_1)} . \tag{19}$$

When $G(x) = 1 - \exp(-\mu x)$ and $c_0(x) = c_0 x$, (19) is rewritten as

$$TC_3(Z_1) = \frac{c_K - \sum_{i=2}^{K}(c_i - c_{i-1})\left[1 - \mathrm{e}^{-\mu(Z_i - Z_1)}\right] - \frac{c_0}{\mu}\mathrm{e}^{-\mu(Z_K - Z_1)}}{1 + \mu Z_1} + \frac{c_0}{\mu} . \tag{20}$$

We find optimal Z_1^* to minimize $C_3(Z_1)$. Differentiating $C_3(Z_1)$ with respect to Z_1 and putting it equal to zero,

$$\mu Z_1 \mathrm{e}^{\mu Z_1}\left[\sum_{i=2}^{K}(c_i - c_{i-1})\mathrm{e}^{-\mu Z_i} - \frac{c_0}{\mu}\mathrm{e}^{-\mu Z_K}\right] = c_1 . \tag{21}$$

Letting denote the left-hand side of Eq. (21) by $Q_3(Z_1)$,

$$\begin{aligned} Q_3(0) &= 0 , \\ Q_3(Z_2) &= \mu Z_2(c_2 - c_1) + \mu Z_2 \mathrm{e}^{\mu Z_2}\left[\sum_{i=3}^{K}(c_i - c_{i-1})\mathrm{e}^{-\mu Z_i} - \frac{c_0}{\mu}\mathrm{e}^{-\mu Z_K}\right] . \end{aligned}$$

Furthermore, because $\exp(-\mu Z_i) > \exp(-\mu Z_K)$, if $(c_K - c_2) > c_0/\mu$, then $Q_3(Z_1)$ increases strictly with Z_1 to $Q_3(Z_2)$. Thus, if $Q_3(Z_2) > c_1$, then there exists a finite and unique Z_1^* $(0 < Z_1^* < Z_2)$ which satisfies (21). When $K = 2$, if $c_2 - c_1 - c_0/\mu > 0$ and

$$\mu Z_2 > \frac{c_1}{c_2 - c_1 - c_0/\mu} ,$$

then a finite Z_1^* exists.

4 Extended Models

We propose the respective extended models of Models 1, 2, and 3, considering a natural disaster and its recovery.

4.1 Model 4

We make assumptions (1), (2'), and (3) of Model 1 and add assumption (4):

(4) Systems are damaged by a natural disaster which occurs independently of steady deteriorations, and its disaster recovery (DR) is done. The occurrence time of a natural disaster has a distribution $F(t)$ and its average DR cost is $c_D(>c_K)$.

The probability that CM is done at time $(j+1)T$ is

$$\sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x)\mathrm{d}G^{(j)}(x) + \sum_{j=0}^{\infty} \overline{F}((j+2)T) \int_0^{Z_1} \left[\int_{Z_1-x}^{Z_K-x} \overline{G}(Z_K - x - y)\mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x), \quad (22)$$

the probability that PM is done at time $(j+1)T$ is

$$\sum_{j=0}^{\infty} \overline{F}((j+2)T) \int_0^{Z_1} \left[\int_{Z_1-x}^{Z_K-x} G(Z_K - x - y)\mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x), \quad (23)$$

and the probability that DR is done at time $(j+1)T$ is

$$\begin{aligned} &\sum_{j=0}^{\infty} F((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x)\mathrm{d}G^{(j)}(x) \\ &+ \sum_{j=0}^{\infty} F((j+2)T) \int_0^{Z_1} \left[\int_{Z_1-x}^{Z_K-x} \overline{G}(Z_K - x - y)\mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x) \\ &+ \sum_{j=0}^{\infty} F((j+2)T) \int_0^{Z_1} \left[\int_{Z_1-x}^{Z_K-x} G(Z_K - x - y)\mathrm{d}G(y) \right] \mathrm{d}G^{(j)}(x) \\ &= \sum_{j=0}^{\infty} F((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x)\mathrm{d}G^{(j)}(x) \\ &+ \sum_{j=0}^{\infty} F((j+2)T) \int_0^{Z_1} [G(Z_K - x) - G(Z_1 - x)]\,\mathrm{d}G^{(j)}(x), \end{aligned} \quad (24)$$

where note that (22) + (23) + (24) = 1. The mean time to maintenance is

$$\begin{aligned}
&\sum_{j=0}^{\infty}[(j+1)T]\overline{F}((j+1)T)\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}[(j+2)T]\overline{F}((j+2)T)\int_0^{Z_1}[G(Z_K-x)-G(Z_1-x)]\,\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}\int_0^{(j+1)T}t\mathrm{d}F(t)\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}\int_0^{(j+2)T}t\mathrm{d}F(t)\int_0^{Z_1}[G(Z_K-x)-G(Z_1-x)]\,\mathrm{d}G^{(j)}(x)\\
&=\int_0^T\overline{F}(t)\mathrm{d}t+\sum_{j=0}^{\infty}\int_{(j+1)T}^{(j+2)T}\overline{F}(t)\mathrm{d}t\int_0^{Z_1}G(Z_K-x)\mathrm{d}G^{(j)}(x)\,, \qquad (25)
\end{aligned}$$

and the total expected cost until maintenance is

$$\begin{aligned}
&c_K\Bigg\{\sum_{j=0}^{\infty}\overline{F}((j+1)T)\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}\overline{F}((j+2)T)\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x}\overline{G}(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\Bigg\}\\
&+c_1\sum_{j=0}^{\infty}\overline{F}((j+2)T)\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x}G(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\\
&+c_D\Bigg\{\sum_{j=0}^{\infty}F((j+1)T)\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+\sum_{j=0}^{\infty}F((j+2)T)\int_0^{Z_1}[G(Z_K-x)-G(Z_1-x)]\,\mathrm{d}G^{(j)}(x)\Bigg\}\\
&=c_K+(c_D-c_K)\sum_{j=0}^{\infty}\Bigg\{F((j+1)T)\int_0^{Z_1}\overline{G}(Z_K-x)\mathrm{d}G^{(j)}(x)\\
&+F((j+2)T)\int_0^{Z_1}[G(Z_K-x)-G(Z_1-x)]\,\mathrm{d}G^{(j)}(x)\Bigg\}\\
&-(c_K-c_1)\sum_{j=0}^{\infty}\overline{F}((j+2)T)\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x}G(Z_K-x-y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x)\,. \qquad (26)
\end{aligned}$$

Thus, the expected cost rate is, from (25) and (26),

$$C_4(Z_1) = \frac{\begin{array}{l} c_K + (c_D - c_K)\sum_{j=0}^{\infty}\left\{F((j+1)T)\int_0^{Z_1}\overline{G}(Z_K - x)\mathrm{d}G^{(j)}(x)\right. \\ \left. +F((j+2)T)\int_0^{Z_1}[G(Z_K - x) - G(Z_1 - x)]\,\mathrm{d}G^{(j)}(x)\right\} \\ -(c_K - c_1)\sum_{j=0}^{\infty}\overline{F}((j+2)T) \\ \times\int_0^{Z_1}\left[\int_{Z_1-x}^{Z_K-x} G(Z_K - x - y)\mathrm{d}G(y)\right]\mathrm{d}G^{(j)}(x) \end{array}}{\int_0^T \overline{F}(t)\mathrm{d}t + \sum_{j=0}^{\infty}\int_{(j+1)T}^{(j+2)T}\overline{F}(t)\mathrm{d}t\int_0^{Z_1} G(Z_K - x)\mathrm{d}G^{(j)}(x)}. \quad (27)$$

When $\overline{F}(t) = 1$, (27) agrees with (9).

Assuming $F(t) = 1 - \exp(-\lambda t)$ and $G(x) = 1 - \exp(-\mu x)$, (27) is rewritten as

$$\frac{C_4(Z_1)}{\lambda} = \frac{\begin{array}{l} c_D - \alpha \mathrm{e}^{\alpha\mu Z_1}\left[(c_D - c_K)\mathrm{e}^{-\mu Z_K} + (c_D - c_1)\alpha\left(\mathrm{e}^{-\mu Z_1} - \mathrm{e}^{-\mu Z_K}\right)\right. \\ \left. -(c_K - c_1)\mu\alpha(Z_K - Z_1)\mathrm{e}^{-\mu Z_K}\right] \end{array}}{1 - \alpha\mathrm{e}^{\alpha\mu Z_1}\left[\alpha\mathrm{e}^{-\mu Z_1} + (1-\alpha)\mathrm{e}^{-\mu Z_K}\right]}, \quad (28)$$

where $\alpha \equiv \exp(-\lambda T)$. We find optimal Z_1^* to minimize $C_4(Z_1)$. Differentiating $C_4(Z_1)$ with respect to Z_1 and putting it equal to zero,

$$\frac{\alpha}{1-\alpha}\mathrm{e}^{-\mu(Z_K - \alpha Z_1)}\left[\mu(Z_K - Z_1)\frac{\mathrm{e}^{(1-\alpha)\mu Z_1} - \alpha}{1 - \mathrm{e}^{-\mu(Z_K - Z_1)}} - (1-\alpha)\right] = \frac{c_1}{c_K - c_1}. \quad (29)$$

Letting $Q_4(Z_1)$ denote the left-hand of (29), and differentiating $Q_4(Z_1)$ with respect to Z_1,

$$\begin{array}{l} \left[-\left(\mathrm{e}^{\mu Z_1} - \alpha\mathrm{e}^{\alpha\mu Z_1}\right) + \mu(Z_K - Z_1)\left(\mathrm{e}^{\mu Z_1} - \alpha^2\mathrm{e}^{\alpha\mu Z_1}\right)\right]\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right] \\ +\mu(Z_K - Z_1)\left(\mathrm{e}^{\mu Z_1} - \alpha\mathrm{e}^{\alpha\mu Z_1}\right)\mathrm{e}^{-\mu(Z_K - Z_1)} - (1-\alpha)\alpha\mathrm{e}^{\alpha\mu Z_1}\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right]^2 \\ > -\left(\mathrm{e}^{\mu Z_1} - \alpha\mathrm{e}^{\alpha\mu Z_1}\right)\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right] + \left(\mathrm{e}^{\mu Z_1} - \alpha^2\mathrm{e}^{\alpha\mu Z_1}\right)\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right]^2 \\ +\left(\mathrm{e}^{\mu Z_1} - \alpha\mathrm{e}^{\alpha\mu Z_1}\right)\mathrm{e}^{-\mu(Z_K - Z_1)}\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right] \\ -(1-\alpha)\alpha\mathrm{e}^{\alpha\mu Z_1}\left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)}\right]^2 = 0. \end{array}$$

Because $x/(1 - \exp(-x))$ is strictly increasing in x from 1 to ∞, $\mu(Z_K - Z_1) > 1 - \exp[-\mu(Z_K - Z_1)]$. Thus, $Q_4(Z_1)$ is strictly increasing in Z_1, and

$$Q_4(0) = \alpha\mathrm{e}^{-\mu Z_K}\left[\frac{\mu Z_K}{1 - \mathrm{e}^{-\mu Z_K}} - 1\right], \quad Q_4(Z_K) = \frac{\alpha}{1-\alpha}\left[1 - \mathrm{e}^{-(1-\alpha)\mu Z_K}\right].$$

Thus, if $Q_4(Z_K) > c_1/(c_K - c_1) > Q_4(0)$, then there exists a finite and unique $Z_1^*(0 < Z_1^* < Z_K)$ which satisfies (29).

4.2 Model 5

We make assumptions (1), (2”), and (3’) of Model 2 and assumption (4) of Model 4. The probability that CM is done at time $(j+1)T$ is

$$\sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x)\,, \tag{30}$$

the probability that PM is done at time $(j+1)T$ is

$$\sum_{i=1}^{K-1} \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}G^{(j)}(x)\,, \tag{31}$$

and the probability that DR is done at time $(j+1)T$ is

$$\sum_{j=0}^{\infty} F((j+1)T) \int_0^{Z_1} \overline{G}(Z_1 - x) \mathrm{d}G^{(j)}(x)\,. \tag{32}$$

The mean time to maintenance is

$$\begin{aligned}
&\sum_{j=0}^{\infty} (j+1)T\overline{F}((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x) \\
&+ \sum_{i=1}^{K-1} \sum_{j=0}^{\infty} (j+1)T\overline{F}((j+1)T) \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}G^{(j)}(x) \\
&+ \sum_{j=0}^{\infty} \int_0^{(j+1)T} t \mathrm{d}F(t) \int_0^{Z_1} \overline{G}(Z_1 - x) \mathrm{d}G^{(j)}(x) = \sum_{j=0}^{\infty} G^{(j)}(Z_1) \int_{jT}^{(j+1)T} \overline{F}(t) \mathrm{d}t\,,
\end{aligned} \tag{33}$$

and the total expected cost until maintenance is

$$\begin{aligned}
&c_K \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x) \\
&+ \sum_{i=1}^{K-1} c_i \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \left[G(Z_{i+1} - x) - G(Z_i - x)\right] \mathrm{d}G^{(j)}(x) \\
&+ c_D \sum_{j=0}^{\infty} F((j+1)T) \int_0^{Z_1} \overline{G}(Z_1 - x) \mathrm{d}G^{(j)}(x)
\end{aligned}$$

$$= c_K - \sum_{i=1}^{K-1}(c_K - c_i)\sum_{j=0}^{\infty}\overline{F}((j+1)T)\int_0^{Z_1}[G(Z_{i+1}-x)-G(Z_i-x)]\mathrm{d}G^{(j)}(x)$$

$$+(c_D - c_K)\sum_{j=0}^{\infty}F((j+1)T)[G^{(j)}(Z_1)-G^{(j+1)}(Z_1)]\,. \tag{34}$$

Thus, the expected cost rate is, from (33) and (34),

$$C_5(Z_1) = \frac{\begin{array}{l} c_K + (c_D - c_K)\sum_{j=0}^{\infty}F((j+1)T)[G^{(j)}(Z_1)-G^{(j+1)}(Z_1)] \\ -\sum_{i=1}^{K-1}(c_K - c_i)\sum_{j=0}^{\infty}\overline{F}((j+1)T) \\ \qquad \times \int_0^{Z_1}[G(Z_{i+1}-x)-G(Z_i-x)]\mathrm{d}G^{(j)}(x) \end{array}}{\sum_{j=0}^{\infty}G^{(j)}(Z_1)\int_{jT}^{(j+1)T}\overline{F}(t)\mathrm{d}t}\,. \tag{35}$$

When $F(t) = 1 - \exp(-\lambda t)$ and $G(x) = 1 - \exp(-\mu x)$, (35) is

$$\frac{C_5(Z_1)}{\lambda} = \frac{c_K - \alpha \mathrm{e}^{-(1-\alpha)\mu Z_1}\sum_{i=2}^{K}(c_i - c_{i-1})\left[1-\mathrm{e}^{-\mu(Z_i-Z_1)}\right]}{1-\alpha\mathrm{e}^{-(1-\alpha)\mu Z_1}} + c_D - c_K, \tag{36}$$

where $\alpha \equiv \exp(-\lambda T)$. We find optimal Z_1^* to minimize $C_5(Z_1)$. Differentiating $C_5(Z_1)$ with Z_1 and putting it equal to zero,

$$\frac{\alpha}{1-\alpha}\left[1-\mathrm{e}^{-(1-\alpha)\mu Z_1}\right]\sum_{i=2}^{K}(c_i - c_{i-1})\mathrm{e}^{-\mu(Z_i-Z_1)} = c_1\,. \tag{37}$$

Letting denote the left-hand side of Eq. (37) by $Q_5(Z_1)$,

$$Q_5(0) = 0\,,$$

$$Q_5(Z_2) = \frac{\alpha}{1-\alpha}\left[1-\mathrm{e}^{-(1-\alpha)\mu Z_2}\right]\left[c_2 - c_1 - \sum_{i=3}^{K}(c_i - c_{i-1})\mathrm{e}^{-\mu(Z_i-Z_2)}\right].$$

Thus, if $Q_5(Z_2) > c_1$, then there exists a finite $Z_1^*\,(0 < Z_1^* < Z_2)$ which satisfies (37).

4.3 Model 6

We make assumptions (1), (2”), and (3”) of Model 3, and assumption (4) of Model 4. Then, the mean time to maintenance is given in (33) and the total expected cost until maintenance is

$$
\begin{aligned}
&[c_K + c_0(Z_K)] \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \overline{G}(Z_K - x) \mathrm{d}G^{(j)}(x) \\
&+ \sum_{i=1}^{K-1} \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \left\{ \int_{Z_i - x}^{Z_{i+1} - x} [c_i + c_0(x+y)] \, \mathrm{d}G(y) \right\} \mathrm{d}G^{(j)}(x) \\
&+ c_D \sum_{j=0}^{\infty} F((j+1)T) \int_0^{Z_1} \overline{G}(Z_1 - x) \mathrm{d}G^{(j)}(x) \\
&= c_K - \sum_{i=1}^{K-1} (c_K - c_i) \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} [G(Z_{i+1} - x) - G(Z_i - x)] \mathrm{d}G^{(j)}(x) \\
&+ (c_D - c_K) \sum_{j=0}^{\infty} F((j+1)T) \left[G^{(j)}(Z_1) - G^{(j+1)}(Z_1) \right] \\
&+ c_0(Z_1) \sum_{j=0}^{\infty} \overline{F}((j+1)T) \left[G^{(j)}(Z_1) - G^{(j+1)}(Z_1) \right] \\
&+ \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \left[\int_{Z_1}^{Z_K} \overline{G}(y - x) \mathrm{d}c_0(y) \right] \mathrm{d}G^{(j)}(x). \qquad (38)
\end{aligned}
$$

Thus, the expected cost rate is, from (33) and (38),

$$
C_6(Z_1) = \frac{\begin{array}{l} c_K + (c_D - c_K) \sum_{j=0}^{\infty} F((j+1)T) \left[G^{(j)}(Z_1) - G^{(j+1)}(Z_1) \right] \\ - \sum_{i=1}^{K-1} (c_K - c_i) \sum_{j=0}^{\infty} \overline{F}((j+1)T) \\ \qquad \times \int_0^{Z_1} [G(Z_{i+1} - x) - G(Z_i - x)] \mathrm{d}G^{(j)}(x) \\ + c_0(Z_1) \sum_{j=0}^{\infty} \overline{F}((j+1)T) \left[G^{(j)}(Z_1) - G^{(j+1)}(Z_1) \right] \\ + \sum_{j=0}^{\infty} \overline{F}((j+1)T) \int_0^{Z_1} \left[\int_{Z_1}^{Z_K} \overline{G}(y - x) \mathrm{d}c_0(y) \right] \mathrm{d}G^{(j)}(x) \end{array}}{\sum_{j=0}^{\infty} G^{(j)}(Z_1) \int_{jT}^{(j+1)T} \overline{F}(t) \mathrm{d}t}. \qquad (39)
$$

When $\overline{F}(t) = 1$, (39) agrees with (20).

When $F(t) = 1 - \exp(-\lambda t)$, $G(x) = 1 - \exp(-\mu x)$, and $c_0(x) = c_0 x$, (39) is

$$
\frac{C_6(Z_1)}{\lambda} = \frac{\begin{array}{l} c_K - \alpha \mathrm{e}^{-(1-\alpha)\mu Z_1} \sum_{i=2}^{K} (c_i - c_{i-1}) \left[1 - \mathrm{e}^{-\mu(Z_i - Z_1)} \right] \\ + c_0 \alpha \mathrm{e}^{-(1-\alpha)\mu Z_1} \left[1 - \mathrm{e}^{-\mu(Z_K - Z_1)} \right] / \mu + c_0 Z_1 \alpha \mathrm{e}^{-(1-\alpha)\mu Z_1} \end{array}}{1 - \alpha \mathrm{e}^{-(1-\alpha)\mu Z_1}} + c_D - c_K. \qquad (40)
$$

We find optimal Z_1^* to minimize $C_6(Z_1)$. Differentiating $C_6(Z_1)$ with Z_1 and putting it equal to zero,

$$\frac{\alpha}{1-\alpha}\left[1-\mathrm{e}^{-(1-\alpha)\mu Z_1}\right]\left[\sum_{i=2}^{K}(c_i-c_{i-1})\mathrm{e}^{-\mu(Z_i-Z_1)}-\frac{c_0}{\mu}\mathrm{e}^{-\mu(Z_K-Z_1)}\right]$$
$$+\frac{c_0}{\mu}\left[\frac{1-\alpha\mathrm{e}^{-(1-\alpha)\mu Z_1}}{1-\alpha}-\mu Z_1\right]=c_1+\frac{c_0}{\mu}. \tag{41}$$

When $c_0 = 0$, (41) agrees with (37).

5 Numerical Calculations

When $T = 1$ and $Z_K = 5$, optimal Z_1^*, and $TC_1(Z_1^*)$ for c_K, c_1, and $1/\mu$ are shown in Table 1. This indicates that When Z_1^* decrease with c_K and $TC_1(Z_1^*)$ increase with c_K, and Z_1^* and $TC(Z_1^*)$ increase with c_1. On the other hand, Z_1^* decrease with $1/\mu$ and $TC(Z_1^*)$ increase with $1/\mu$.

When $T = 1$, $Z_K = 5$, $c_i = ai$, and $Z_i = i$, optimal Z_1^* and $TC_2(Z_1^*)$ for c_K, a, and $1/\mu$ are shown in Table 2. This indicates that Z_1^* decrease with c_K and $TC_2(Z_1^*)$ increase with c_K, and Z_1^* and $TC_2(Z_1^*)$ increase with a. Z_1^* decrease with $1/\mu$ and $TC_2(Z_1^*)$ increase with $1/\mu$. These tendencies are same as ones in Model 1.

Table 3 presents optimal Z_1^* and $TC_3(Z_1^*)$ for c_K, a, and $1/\mu$ when $T = 1$, $Z_K = 5$, $c_i = ai$, and $Z_i = i$. This indicates that Z_1^* and $TC_3(Z_1^*)$ increase with c_0. In this illustration, when c_K, a, and $1/\mu$ are changed, variations of Z_1^* and $TC_3(Z_1^*)$ are similar ones in Table 2.

When $T = 1$, $Z_K = 5$, and $c_D = 100$ for c_K, c_1, $1/\mu$, and λ, optimal Z_1^* and $C_4(Z_1^*)$ are shown in Table 4. This indicates that Z_1^* and $C_4(Z_1^*)$ decrease with $1/\lambda$. When c_K, c_1, and $1/\mu$ are changed, variations of Z_1^* and $C_4(Z_1^*)$ are similar ones in Table 1.

Table 1 Optimal Z_1^* and $TC_1(Z_1^*)$ when $T = 1$, $Z_K = 5$

c_K	c_1	$1/\mu$	Z_1^*	$TC_1(Z_1^*)$
25	1	1	0.230	0.979
30	1	1	0.123	1.087
35	1	1	0.040	1.191
25	2	1	0.700	1.360
25	3	1	1.031	1.682
25	1	1/1.5	1.097	0.404
25	1	1/2	1.740	0.231

Table 5 shows the numerical examples in Model 5 for c_K, c_1, $1/\mu$, a, and $1/\lambda$ when $T = 1$, $Z_K = 5$, $c_i = ai$, $Z_i = i$, and $c_D = 100$. This indicates that Z_1^* and $C_5(Z_1^*)$ decrease with $1/\lambda$. When c_K and $1/\mu$ are changed, variations of Z_1^* and $C_5(Z_1^*)$ are similar ones in Tables 2 and 3.

Table 6 presents optimal Z_1^* and $C_6(Z_1^*)$ for c_K, c_0, $1/\mu$, a, and $1/\lambda$, when $T = 1$, $Z_K = 5$, $c_i = ai$, $Z_i = i$, and $c_D = 100$. When c_K, a, c_0, $1/\mu$, and $1/\lambda$ are changed, variations of Z_1^* and $C_6(Z_1^*)$ are similar ones in Tables 2, 3, and 5.

$TC_3(Z_1^*)$ and $C_6(Z_1^*)$ are larger than $TC_2(Z_1^*)$ and $C_5(Z_1^*)$, respectively, because Models 3 and 6 introduce modified costs. $C_4(Z_1^*)$, $C_5(Z_1^*)$ and $C_6(Z_1^*)$ are larger than $TC_1(Z_1^*)$, $TC_2(Z_1^*)$ and $TC_3(Z_1^*)$, respectively, because Models 4–6 consider disaster recovery costs. Optimal Z_1^* of Models 3 and 6 are greater than Z_1^* of Models 2 and 5, respectively, and Z_1^* of Models 5 and 6 are greater than those of Models 2 and 3, respectively, in this numerical illustrations. It means that the risk of CM is low when Z_1^* of model is large. Although most Z_1^* of Models 1 and 4 are smaller than those of Models 2 and 5, Z_1^* of Models 1 and 4 are greater than those of Models 2 and 5 when $1/\mu = 1/2$. It means that although PM delays are more risk than the multiple maintenance costs models in most cases, there exist some exceptions.

Table 2 Optimal Z_1^* and $TC_2(Z_1^*)$ when $T = 1$, $Z_K = 5$, $c_i = ai$, and $Z_i = i$

c_K	a	$1/\mu$	Z_1^*	$TC_2(Z_1^*)$
25	1	1	1.444	1.385
30	1	1	1.387	1.442
35	1	1	1.335	1.498
25	2	1	1.625	2.462
25	3	1	1.702	3.526
25	1	1/1.5	1.602	0.832
25	1	1/2	1.657	0.604

Table 3 Optimal Z_1^* and $TC_3(Z_1^*)$ when $T = 1$, $Z_K = 5$, $c_i = ai$, and $Z_i = i$

c_K	a	c_0	$1/\mu$	Z_1^*	$TC_3(Z_1^*)$
25	1	1	1	1.457	2.373
30	1	1	1	1.398	2.431
35	1	1	1	1.345	2.487
25	2	1	1	1.634	3.449
25	3	1	1	1.708	4.512
25	1	2	1	1.469	3.361
25	1	3	1	1.482	4.349
25	1	1	1/1.5	1.622	1.489
25	1	1	1/2	1.661	1.102

Table 4 Optimal Z_1^* and $C_4(Z_1^*)$ when $T = 1$, $Z_K = 5$, and $c_D = 100$

c_K	c_1	$1/\mu$	$1/\lambda$	Z_1^*	$C_4(Z_1^*)$
25	1	1	10	0.303	10.887
30	1	1	10	0.188	10.988
35	1	1	10	0.097	11.085
25	2	1	10	0.811	11.211
25	3	1	10	1.170	11.478
25	1	1/1.5	10	1.214	10.344
25	1	1/2	10	1.884	10.180
25	1	1	50	0.244	2.960
25	1	1	100	0.236	1.969

Table 5 Optimal Z_1^* and $C_5(Z_1^*)$ when when $T = 1$, $Z_K = 5$, $c_D = 100$, $c_i = ai$, and $Z_i = i$

c_K	a	$1/\mu$	$1/\lambda$	Z_1^*	$C_5(Z_1^*)$
25	1	1	10	1.548	11.260
30	1	1	10	1.487	11.316
35	1	1	10	1.433	11.369
25	2	1	10	1.739	12.223
25	3	1	10	1.820	13.173
25	1	1/1.5	10	1.725	10.716
25	1	1/2	10	1.762	10.502
25	1	1	50	1.465	3.359
25	1	1	100	1.455	2.372

Table 6 Optimal Z_1^* and $C_6(Z_1^*)$ when $T = 1$, $Z_K = 5$, $c_D = 100$, $c_i = ai$, and $Z_i = i$

c_K	a	c_0	$1/\mu$	$1/\lambda$	Z_1^*	$C_6(Z_1^*)$
25	1	1	1	10	1.640	12.155
30	1	1	1	10	1.572	12.213
35	1	1	1	10	1.512	12.270
25	2	1	1	10	1.739	13.110
25	3	1	1	10	1.640	14.056
25	1	2	1	10	1.737	13.045
25	1	3	1	10	1.838	13.931
25	1	1	1/1.5	10	1.812	11.292
25	1	1	1/2	10	1.843	10.915
25	1	1	1	50	1.492	4.329
25	1	1	1	100	1.475	3.351

6 Conclusion

The maintenance of social infrastructures have some remarkable characteristics which are different from ordinary systems. PM delays occur inevitably because the annual budget of local government has a limitation. Because social infrastructures are huge and their PM cost ranges is extremely wide due to degradation levels, multiple degradation levels and modified costs which depend on damage levels should be introduced. Natural disasters should be considered when PM of social infrastructures are planned because they have huge scales and have been operated for decades.

In this chapter, we have considered optimal maintenance policies considering delay, multiple degradation levels, modified costs which depend on damage levels, and disaster recovery. The six maintenance policies have been proposed using cumulative damage model. Model 1 has considered delayed maintenance, Model 2 has regarded multiple degradation levels, and Model 3 has introduced modified costs which depend on damage levels. Models 4, 5, and 6 are extended models of Models 1, 2, and 3 in which the disaster recovery has been supposed. The expected cost rates have been obtained and optimal damage levels Z_1^* which minimize them have been discussed analytically and numerically. Although most of numerical results are acceptable intuitively, it is of interest that Z_1^* of Models 1 and 4 are greater than those of Models 2 and 5 when $1/\mu = 1/2$ in this illustration. These policies would be greatly useful when actual PMs of social infrastructures would be planned, referring to optimal policies which minimize the total cost.

References

1. Ministry of Land, Infrastructure, Transport and Tourism (2017) White paper on land, infrastructure, transport and tourism in Japan, p 114
2. Friesz T, Fernandez J (1979) A model of optimal transport maintenance with demand responsiveness. Transp Res Part B: Methodol 13(4):317–339
3. Markow M, Balta W (1985) Optimal rehabilitation frequencies for highway pavements. Transp Res Rec 1035:31–43
4. Tsunokawa K, Schofer J (1994) Trend curve optimal control model for highway pavement maintenance: case study and evaluation. Transp Res Part A 28(2):151–166
5. Li Y, Madanat S (2002) A steady-state solution for the optimal pavement resurfacing problem. Transp Res Part A 36(6):525–535
6. Ouyang Y, Madanat S (2004) Optimal scheduling of rehabilitation activities for multiple pavement facilities: exact and approximate solutions. Transp Res Part A 38(5):347–365
7. Ouyang Y, Madanat S (2006) An analytical solution for the finite-horizon pavement resurfacing planning problem. Transp Res Part B 40(9):767–778
8. Kobayashi K, Eguchi M, Oi A, Aoki K, Kaito K, Matsumura Y (2012) The optimal repair and replacement model of pavement structure. J Jpn Soc Civ Eng E1 68(2):54–68
9. Kaito K, Yasuda K, Kobayashi K, Owada K (2005) Optimal maintenance strategies of bridge components with an average cost minimizing principles. Proc J Jpn Soc Civ Eng (801):83–96
10. Sakai R, Yuzo O, Otsu H (2004) The study of the maintenance management model of the tunnel structure using stochastic process. In: 39th Proceedings of the Japan national conference on geotechnical engineering, 2-2, pp 1711–1712

11. Obama K, Kaito K, Aoki K, Kobayashi K, Fukuda T (2012) The optimal scrapping and maintenance model of infrastructure considering deterioration process. J Jpn Soc Civ Eng F4 68(3):141–156
12. Tsuda Y, Kaito K, Aoki K, Kobayashi K (2005) Estimating Markovian transition probabilities for bridge deterioration forecasting. J Jpn Soc Civ Eng 801/I-73:69–82
13. Ito K, Higuchi Y, Nakagawa T (2018) Optimal maintenance policy of coastal protection systems. In: Conference proceedings, 24th ISSAT international conference reliability and quality in design, Toronto, Ontario, Canada, 2–4 August 2018
14. Ito K, Nakagawa T (2018) Optimal maintenance policies of social infrastructures. IEICE Tech Rep 118(365):13–16, R2018-45
15. Kishida T, Ito K, Higuchi Y, Nakagawa T (2018) Optimal maintenance policy of coastal protection systems. In: Conference proceedings, 25th ISSAT international conference reliability and quality in design, Las Vegas, Nevada, USA, 2–4 August 2018
16. Nakagawa T (2007) Shock and damage models in reliability theory. Springer, London

Takumi Kishida received the B.S. degree from Tottori University in 2019. He is now a graduate student of Tottori University. His research interests include maintenance policies of social Infrastructures.

Kodo Ito received the D.E. degree from Aichi Institute of Technology in 1999. He worked as an engineer at Mitsubishi Heavy Industries, LTD. from 1985 to 2017. He is now a Professor with Tottori University, Japan, and Visiting Professor with Fukushima University, Japan. His research interests are maintenance policies of huge and complicated systems such as aircrafts and social infrastructures.

Higuchi Yoshiyuki received the D.E. degree from Nagaoka University of Technology (NUT) in 2000. He worked as an associate professor at NUT from 2000 to 2004 and Fukushima University, Japan from 2004 to 2013. Now, he is a professor at Fukushima University, Japan. His research interests are system modeling, system simulation analysis, optimal design, operation, and maintenance in social infrastructures.

Toshio Nakagawa received B.S.E. and M.S. degrees from Nagoya Institute of Technology in 1965 and 1967, Japan, respectively; and a Ph.D. degree from Kyoto University in 1977, Japan. He worked as a Research Associate at Syracuse University for two years from 1972 to 1973. He is now a Honorary Professor with Aichi Institute of Technology, Japan. He has published 5 books from Springer, and about 200 journal papers. His research interests are optimization problems in operations research and management science, and analysis for stochastic and computer systems in reliability and maintenance theory.

Optimal Checkpoint Intervals, Schemes and Structures for Computing Modules

Kenichiro Naruse and Toshio Nakagawa

Abstract This chapter takes up a high-reliability computing system with redundancy techniques and recovery methods to prevent failure occurrences. We adopt duplex and majority decision modular systems as redundancy techniques, and apply them to two checkpoint schemes in which their respective interval times are periodic and random. Introducing overheads for recovery and checkpoint, we obtain the mean execution times until the process succeeds, and derive optimal checkpoint times to minimize them. Furthermore, we compare a duplex system and a majority decision system, and discuss analytically and numerically which system is better.

Keywords Recovery method · Redundancy technique · Checkpoint schemes · Random interval · Duplex system · Majority system

1 Introduction

Most computer systems have been required for operating normally and effectively as communication and information systems have been developed rapidly and complicated greatly. However, some errors occur due to noises, human errors, hardware faults, computer viruses, and cyber-attacks. Lastly, these errors might have become faults and incur system failures. Such failures have sometimes caused heavy damage to human society, which has fallen into general disorder. To prevent such faults, various kinds of fault-tolerant techniques such as the redundancy and configuration of systems have been provided [1, 8, 21]. The high-reliability and effective performance of real systems can be achieved by several kinds of fault-tolerant techniques.

K. Naruse (✉)
Nagoya Sangyo University, 3255-5 Yamanota Araicho, Owariasahi, Aichi, Japan
e-mail: naruse@nagoya-su.ac.jp

T. Nakagawa
Department of Business Administration, Aichi Institute of Technology, Toyota, Aichi, Japan
e-mail: toshi-nakagawa@aitech.ac.jp

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_16

Partial data loss and operational errors in computer systems are generally called *error* and *fault*. Failure indicates that faults are recognized on exterior systems. Three different techniques of decreasing the possibility of fault occurrences can be used in actual fields [8]: *Fault avoidance* is to prevent fault occurrences by improving the qualities of structure parts and placing well surrounding, *Fault masking* is to prevent faults by error correction codes and majority voting, and *Fault tolerance* is that systems continue to function correctly in the presence of hardware failures and software errors.

Some faults due to operational errors may be detected after some time has passed and a system consistency may be lost. For protecting such serious events, we should restore a consistent state just before failing occurrences by some recovery methods. The method which takes copies of a normal state of the system is called *checkpoint*. When faults have occurred, the process goes back to the nearest checkpoint by roll-back operation [10], and its retry is made, using the copies of a consistent state at checkpoint times.

2 Failure Detection and Recovery Methods

It is important to know useful methods for failure detection because most systems have become larger and more complex recently. Several methods to detect failures have been proposed: O'Connor [17] surveyed widely the methods related to tests for electronic circuits. Lala [7] summarized fault-tolerant design methods with the self-checking of digital circuits. Such methods of self-checking which involve fault-secure and self-testing are required for designing high reliable systems. Fault-secure means that a failed system outputs codes except for an assumed output code space, and self-testing means that a failed system outputs codes except for an assumed code space for at least one input code. Another method for systems such as digital circuits is the comparison-checking with outputs of duplex modular systems: Two modules execute the same process and compare two states at checkpoint times. If two states of each module do not match with each other, this means system failure. One extending system is a majority decision system, i.e., an $(n + 1)$-out-of-$(2n + 1)$ system as an error masking system. If $(n + 1)$ or more states of $(2n + 1)$ modules match, the system is correct. On the other hand, when the logical consistency is lost by failures, the following two operations of process recoveries are performed: One is forward recovery which keeps running forward in a fault status without backward. Such a method has been applied to real-time systems as weather satellite's picture and to forward the voice of the IP telephony system. The other is backward recovery which goes back before failure occurrence. Typical three methods of forward and backward recoveries have been generally used [1, 6, 8, 13]:

i. Retry Recovery Method
   Retry recovery method is a very popular and easy one: If faults occur, the procedure of retry is made immediately. But if the fault changes an original data

into another one, the retry fails. This method is used for hard disk read, memory read and so on.

ii. Checkpoint Recovery Method
This is the most general method of the backward recovery system. This method records system states which need to run continually the process at suitable intervals. If faults are detected at some checkpoint, the process goes back to the latest checkpoint. It is important to determine the interval times of checkpoints. If we generate checkpoints at short intervals, their overheads are large, and the system performance becomes low. But, if some errors are detected in the process, the process goes back to the nearest checkpoint and is re-executed. In this case, the re-execution time is small. On the other hand, if we generate checkpoints at long intervals, their overheads are small and the system performance becomes high. But, if some errors are detected in the process, the process goes back to the far checkpoint and is re-executed. In this case, the re-execution time is large. Therefore, it is one kind of trade-off problems to generate checkpoint intervals.

iii. Journal Recovery Method
Journal recovery method is an easy one, but, it needs a longer time than the above two methods at failures. This method records an initial state and records all of the transactions about changing data. If failures occur, the process is re-executed from initial data of all transactions, which needs a longer time than the above two methods.

High-reliability systems can be achieved by redundancy techniques. A classical standard problem is to determine how reliability can be improved by using redundant modules. The results of various redundant systems with the repair were summarized as repairman problem, and optimization problems of redundancy and allocation subject to some constraints were solved and qualitative relationships for multicomponent structures obtained [2]. Furthermore, some useful expressions of reliability measures of many redundant systems were shown [3, 23]. The fundamentals and applications of system reliabilities and reliability optimizations in system design were well described [5]. Various optimization problems of combinatorial reliabilities with multiple constraints for different system structures were considered [22] and their computational techniques were surveyed [9].

Redundancy techniques of a system for improving reliability and achieving fault tolerance are classified commonly into the following types [1, 8, 13]:

(1) Hardware Redundancy
This can be classified into three kinds of redundancies.

(a) Static hardware redundancy is a fault-masking technique in which the effects of faults are essentially hidden from the system with no specific indication of their occurrences and existing faults are not removed. A typical example is a triple modular redundancy.
(b) Dynamic hardware redundancy is a fault tolerance technique in which the system continues to function by detection and removing faults, replacing

faulty units and making reconfiguration. Typical examples are a standby sparing system and a graceful degrading system [6].

(c) Hybrid hardware redundancy is a combination of the advantages of static and dynamic hardware redundancies.

(2) Software Redundancy
This is to use extra codes, small routines or possibly complete programs to check the correctness or consistency of results produced by software. Typical examples are N-version programming and Ad-Hoc technique.

(3) Information Redundancy
This adds redundant information to data to allow fault detection, fault-masking and fault tolerance. Typical examples are error-detecting codes such as parity codes and signatures, and a watchdog processor.

(4) Time Redundancy
This is the several repetitions of given computation and the comparisons of results. This is used to detect transient of intermittent faults to mask faults and to recover the system. Typical examples are retries and checkpoint schemes.

Redundancies (1), (2) and (3) are also called *Space Redundancy* because high reliability is attained by providing multiple resources of hardware and software. In this chapter, we take up checkpoint schemes for redundant systems as recovery methods.

Suppose that we have to complete the process of tasks with finite execution times. Modules are elements such as logical circuits or processors which execute certain lumped parts of tasks. For such systems, we consider checkpoint schemes of error detection and masking by redundancy, and propose their modified and extended models.

Several studies of deciding optimal checkpoint frequencies have been done: The performance and reliability of redundant modular systems [10, 20] and the performance of checkpoint schemes with task duplication [25, 26] were evaluated. Furthermore, the optimal instruction-retry period that minimizes the probability of dynamic failure by a triple modular controller was derived [4]. Therefore, the evaluation models with finite checkpoints and bounded rollbacks were discussed [18] and the checkpointing schemes for a set of multiple tasks in real-time systems were investigated [24].

Section 3 adopts multiple modular redundant systems as the recovery methods of error detection and error masking, and derives optimal checkpoint times. Section 4 adopts tasks with random processing times which are executed successively and two types of checkpoints are placed at the end of tasks. Three checkpoint schemes are considered and compared numerically. Finally, it is assumed in Sect. 5 that when the number of computing modules consists of 6 modules, we discuss which structure is optimal for a duplex structure or a majority structure. This chapter is written mainly based on [10, 12, 14–16].

3 Periodic Checkpoint Models

We introduce the redundant techniques of error detection and error masking on finite process execution. First, an error detection of the process can be made by two independent modules where they compare two states at periodic checkpoint times. If their states do not match with each other, we go back to the previous checkpoint and make a retrial of the process. Secondly, a majority decision system with multiple modules is adopted as the technique of an error masking and the state is decided by its majority. In this case, we determine numerically which majority decision system is optimal.

We consider a duplex modular system and a majority decision system with $(2n + 1)$ $(n = 1, 2, \ldots)$ modules of an error masking system, compute the mean times to completion of the process and compare them numerically.

(1) Duplex Modular System

Suppose that S $(0 < S < \infty)$ is a native execution time of the process which does not include the overheads of retries and checkpoint generations. Then, we divide S equally into N $(N = 1, 2, \ldots)$ parts and create checkpoints at periodic times kT $(k = 1, 2, \ldots, N - 1)$, where $S = NT$ in Fig. 1.

To detect errors, we provide two independent modules where they compare two states at periodic times kT $(k = 1, 2, \ldots, N)$. If two states match with each other, two processes are correct and go forward. However, if two states do not match, it is judged that some errors have occurred. Then, we make a rollback operation to the previous checkpoint and retry of the process. The process completes when all of two processes are successful for N intervals.

Let introduce the overhead C_1 for the comparison of two states. We neglect any failures of the system caused by faults with common-mode to make clear error detections of the process. Furthermore, it is assumed that some error of one process occurs at constant rate λ $(0 < \lambda < \infty)$, i.e., the probability that any errors do not occur during $(0, t]$ is given by $\mathrm{e}^{-\lambda t}$. Thus, the probability that two processes have no error during $(0, t]$ is $\overline{F}_1(t) = \mathrm{e}^{-2\lambda t}$ [19].

The mean execution time $l_1(N)$ to completion of the process is the summation of the processing times and the overhead C_1 of comparison of two processes. When some error has been detected at a checkpoint, two processes are rolled back to the previous checkpoint. Then, the mean execution time $\widetilde{l}_1(1)$ of the process for one checkpoint interval $(0, T]$ is given by a renewal equation:

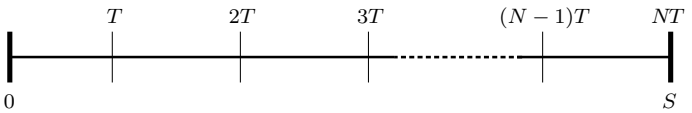


Fig. 1 Periodic checkpoint intervals

$$\widetilde{l}_1(1) = T + C_1 + \widetilde{l}_1(1)\left(1 - \mathrm{e}^{-2\lambda T}\right), \tag{1}$$

and solving it for $\widetilde{l}_1(1)$,

$$\widetilde{l}_1(1) = (T + C_1)\,\mathrm{e}^{2\lambda T}. \tag{2}$$

Thus, the mean execution time to completion of the process is

$$l_1(N) \equiv N\widetilde{l}_1(1) = N(T + C_1)\,\mathrm{e}^{2\lambda T} = (S + NC_1)\,\mathrm{e}^{2\lambda S/N} \ (N = 1, 2, \ldots). \tag{3}$$

We find optimal number N_1^* to minimize $l_1(N)$ for a specified S. Noting that $\lim_{N\to\infty} l_1(N) = \infty$, a finite N_1^* exists. Setting $T \equiv S/N$ in (3) and rewriting it by the function of $T\,(0 < T \le S)$,

$$l_1(T) = S\left(1 + \frac{C_1}{T}\right)\mathrm{e}^{2\lambda T}. \tag{4}$$

Differentiating $l_1(T)$ with respect to T and setting it equal to zero,

$$T^2 + C_1 T - \frac{C_1}{2\lambda} = 0. \tag{5}$$

Solving it with T,

$$\widetilde{T}_1 = \frac{C_1}{2}\left[\sqrt{1 + \frac{2}{\lambda C_1}} - 1\right]. \tag{6}$$

Therefore, we have the following optimal interval number N_1^* [11]:

1. If $\widetilde{T}_1 < S$, we put $[S/\widetilde{T}_1] \equiv N$, where $[x]$ denotes the largest integer contained in x, and compute $l_1(N)$ and $l_1(N+1)$ from (3). If $l_1(N) \le l_1(N+1)$ then $N_1^* = N$ and $T_1^* = S/N_1^*$, and conversely, if $l_1(N+1) < l_1(N)$ then $N_1^* = N + 1$.
2. If $\widetilde{T}_1 \ge S$, i.e., $N_1^* = 1$ and we should make no checkpoint until time S, and the mean execution time is $l_1(S) = (S + C_1)\mathrm{e}^{2\lambda S}$.

Note that $\widetilde{T}_1$ in (6) does not depend on S. Thus, if S is very large, is changed greatly or is unclear, then we might adopt $\widetilde{T}_1$ as an approximate checkpoint time.

Furthermore, the mean execution time per one checkpoint interval is

$$\widetilde{l}_1(T) \equiv \frac{l_1(T)}{T} = \left(1 + \frac{C_1}{T}\right)\mathrm{e}^{2\lambda T}. \tag{7}$$

Thus, optimal time to minimize $\widetilde{l}_1(T)$ also agrees with $\widetilde{T}_1$ in (6).

Example 1 Table 1 presents $\lambda\widetilde{T}_1$ in (6), optimal N_1^*, λT_1^* and $\lambda l_1(N_1^*)$ for λC_1 when $\lambda S = 0.1$. This indicates that N_1^* decrease with λC_1 and T_1^* increase with λC_1. For example, when $\lambda = 0.01$ (1/s), $C_1 = 0.1$ (1/s) and $S = 10.0$ (s), $N_1^* = 5$, $T_1^* = S/5 = 2.0$ (s), and $l_1(5) = 10.929$ (s), which is longer about 9.3% than S. □

Table 1 Optimal N_1^*, λT_1^* and mean execution time $\lambda l_1(N_1^*)$ when $\lambda S = 0.1$

$\lambda C_1 \times 10^3$	$\lambda \widetilde{T}_1 \times 10^2$	N_1^*	$\lambda l_1(N_1^*) \times 10^2$	$\lambda T_1^* \times 10^2$
0.1	0.702	14	10.286	0.714
0.2	0.990	10	10.406	1.000
0.3	1.210	8	10.499	1.250
0.4	1.394	7	10.578	1.429
0.5	1.556	6	10.650	1.667
1.0	2.187	5	10.929	2.000
1.5	2.665	4	11.143	2.500
2.0	3.064	3	11.331	3.333
3.0	3.726	3	11.715	3.333
4.0	4.277	2	11.936	5.000
5.0	4.756	2	12.157	5.000
10.0	6.589	2	13.435	5.000
20.0	9.050	1	14.657	10.000

(2) Triple Majority Decision System

Consider a majority decision system with three modules, i.e., 2-out-of-3 system as an error masking system. If more than two states of three modules match equally, then the process in this interval is correct, i.e., the system can mask a single module error. Then, the probability that the process is correct during each interval $((k-1)T, kT]$ is

$$\overline{F}_2(T) = \mathrm{e}^{-3\lambda T} + 3\mathrm{e}^{-2\lambda T}\left(1 - \mathrm{e}^{-\lambda T}\right) = 3\mathrm{e}^{-2\lambda T} - 2\mathrm{e}^{-3\lambda T}. \tag{8}$$

Let C_2 be the overhead for the comparison of a majority decision with 3 modules. By the similar method of obtaining (3), the mean execution time to completion of the process is

$$l_2(N) = \frac{N(T + C_2)}{\overline{F}_2(T)} = \frac{S + NC_2}{3\mathrm{e}^{-2\lambda T} - 2\mathrm{e}^{-3\lambda T}} \qquad (N = 1, 2, \ldots). \tag{9}$$

Setting $T \equiv S/N$ in (9),

$$l_2(T) = \frac{S(1 + C_2/T)}{3\mathrm{e}^{-2\lambda T} - 2\mathrm{e}^{-3\lambda T}}. \tag{10}$$

Differentiating $l_2(T)$ with respect to T and setting it equal to zero,

$$\left(\mathrm{e}^{\lambda T} - 1\right)\left(T^2 + C_2 T - \frac{C_2}{2\lambda}\right) = \frac{C_2}{6\lambda}, \tag{11}$$

Table 2 Optimal N_2^*, λT_2^* and mean execution time $\lambda l_2(N_2^*)$ when $\lambda S = 0.1$

$\lambda C_2 \times 10^3$	$\lambda \widetilde{T}_2 \times 10^2$	N_2^*	$\lambda l_2(N_2^*) \times 10^2$	$\lambda T_2^* \times 10^2$
0.1	2.605	3	10.050	3.333
0.2	3.298	3	10.091	3.333
0.3	3.787	2	10.101	5.000
0.4	4.178	2	10.129	5.000
0.5	4.510	2	10.158	5.000
1.0	5.721	1	10.191	10.000
1.5	6.579	1	10.270	10.000
2.0	7.265	1	10.345	10.000
3.0	8.358	1	10.492	10.000
4.0	9.232	1	10.634	10.000
5.0	9.974	1	10.773	10.000
10.0	12.675	1	11.448	10.000
20.0	16.087	1	12.762	10.000

whose left-hand side increases strictly with T from 0 to ∞. Thus, there exists a finite and unique $\widetilde{T}_2 (0 < \widetilde{T}_2 < \infty)$ which satisfies (11).

Therefore, using a similar method of 1. and 2., we get optimal N_2^* and T_2^* which minimize $l_2(N)$ in (9). In particular, when $C_1 = C_2$, it is easily shown from (5) and (11) that $\widetilde{T}_2 > \widetilde{T}_1$, i.e., $N_2^* \leq N_1^*$.

Example 2 Table 2 presents $\lambda \widetilde{T}_2$ in (11), optimal N_2^*, λT_2^* and $\lambda l_2(N_2^*)$ for λC_2 when $\lambda S = 0.1$. For example, when $\lambda = 0.01$ (1/s), $C_2 = 0.1$ (s) and $S = 10$ (s), $N_2^* = 1$, $T_2^* = 10.000$ (s), and $l_2(2) = 10.191$ (s), which is longer about 1.2% than S. It can be easily seen that optimal $N_i^* (i = 1, 2)$ decrease as overhead C_i increase in both Tables 1 and 2. When $C_1 = C_2$, $N_1^* \geq N_2^*$ and $l_1(N_1^*) \geq l_2(N_2^*)$, as shown above theoretically. □

(3) Majority Decision System

Consider a majority decision system with $(2n + 1)$ modules as an error masking system with constant error rate λ, i.e., an $(n + 1)$-out-of-$(2n + 1)$ system $(n = 1, 2, \ldots)$. If more than $(n + 1)$ states of $(2n + 1)$ modules match equally, the process in this interval is correct. Then, the probability that the process is correct during $((k - 1)T, kT]$ is

$$\overline{F}_{n+1}(T) = \sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \left(\mathrm{e}^{-\lambda T}\right)^j \left(1 - \mathrm{e}^{-\lambda T}\right)^{2n+1-j}. \tag{12}$$

Let C_{n+1} be the overhead for the comparison of a majority decision with $(2n + 1)$ modules. By a similar method of obtaining (9), the mean execution time to completion of the process is

Table 3 Optimal N^*_{n+1} and mean execution time $\lambda l_{n+1}(N^*_{n+1})$ when $\lambda S = 0.1$

n	$\lambda C_1 = 0.05 \times 10^{-3}$		$\lambda C_1 = 0.1 \times 10^{-3}$		$\lambda C_1 = 0.5 \times 10^{-3}$	
	N^*_{n+1}	$\lambda l_{n+1}(N^*_{n+1}) \times 10^2$	N^*_{n+1}	$\lambda l_{n+1}(N^*_{n+1}) \times 10^2$	N^*_{n+1}	$\lambda l_{n+1}(N^*_{n+1}) \times 10^2$
1	3	10.077	3	10.122	2	10.372
2	2	10.111	1	10.176	1	10.579
3	1	10.128	1	10.233	1	11.075
4	1	10.187	1	10.367	1	11.808

$$l_{n+1}(N) = \frac{N(T + C_{n+1})}{\overline{F}_{n+1}(T)} = \frac{S + NC_{n+1}}{\sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \mathrm{e}^{-j\lambda T} \left(1 - \mathrm{e}^{-\lambda T}\right)^{2n+1-j}} \quad (N = 1, 2, \ldots), \quad (13)$$

which agrees with (9) when $n = 1$.

Next, we consider the problem of which majority system is optimal, when the overhead for an $(n + 1)$-out-of-$(2n + 1)$ system is given by

$$C_{n+1} \equiv \binom{2n+1}{2} C_1 \quad (n = 1, 2, \ldots),$$

because we select two states and compare them from each of $(2n + 1)$ modules. For example, $C_2 = 3C_1$, $C_3 = 10C_1, \ldots$.

Example 3 Table 3 presents optimal N^*_{n+1} and the mean execution time $\lambda l_{n+1}(N^*_{n+1}) \times 10^2$ for n and λC_1 when $\lambda S = 0.1$. For example, when $\lambda C_1 = 0.5 \times 10^{-3}$, optimal number is $N^*_2 = 2$ and $\lambda l_2(2) = 10.372 \times 10^{-2}$, which is the smallest among these systems, i.e., a 2-out-of-3 system with $N^*_2 = 2$ is optimal. The mean times for $n = 1, 2$ when $\lambda C_1 = 0.5 \times 10^{-3}$ are smaller than 10.650×10^{-2} in Table 1 for a duplex modular system. □

4 Random Checkpoint Models

Suppose that we have to execute the successive tasks with a processing time Y_k $(k = 1, 2, \ldots)$ in Fig. 2. A duplex modular system of error detection for the processing of each task is adopted. Then, setting two types of checkpoints; compare-and-store checkpoint (CSCP) and compare-checkpoint (CCP), we consider the following three checkpoint schemes:

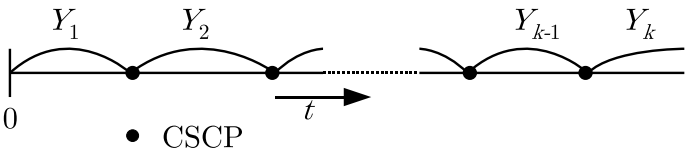


Fig. 2 Task execution for Scheme 1

1. CSCP is placed at each end of tasks.
2. CSCP is placed at the Nth ($N = 1, 2, \ldots$) end of tasks.
3. CCP is placed at each end of tasks and CSCP is placed at the Nth end of tasks.

4.1 Duplex Modular System

Suppose that each task has a processing time Y_k ($k = 1, 2, \ldots$) with an identical distribution $G(t) \equiv \Pr\{Y_k \leq t\}$ and finite mean $1/\theta \equiv \int_0^\infty \overline{G}(t)\,\mathrm{d}t < \infty$, and is executed successively. To detect errors, we provide two independent modules where they compare two states at checkpoint times. In addition, some error occurs at constant rate $\lambda(\lambda > 0)$, i.e., the probability that two modules have no error during $(0, t]$ is $\mathrm{e}^{-2\lambda t}$.

(1) Scheme 1

CSCP is placed at each end of task k in Fig. 2: When two states of modules match with each other at the end of task k, the process of task k is correct and its state is stored. In this case, two modules go forward and execute task $k + 1$. However, when two states do not match, it is judged that some errors have occurred. Then, two modules go back and make the retry of task k again.

Let C be the overhead for the comparison of two states and C_S be the overhead for their store. Then, the mean execution time of the process of task k is given by a renewal equation:

$$\widetilde{l}_1(1) = C + \int_0^\infty \left\{\mathrm{e}^{-2\lambda t}(C_S + t) + \left(1 - \mathrm{e}^{-2\lambda t}\right)\left[t + \widetilde{l}_1(1)\right]\right\}\mathrm{d}G(t), \tag{14}$$

and solving it for $\widetilde{l}_1(1)$,

$$\widetilde{l}_1(1) = \frac{C + 1/\theta}{G^*(2\lambda)} + C_S,$$

where $G^*(s)$ is the Laplace-Stieltjes (LS) transform of $G(t)$, i.e., $G^*(s) \equiv \int_0^\infty \mathrm{e}^{-st}\mathrm{d}G(t)$ for $s > 0$. Therefore, the mean execution time per one task is

$$l_1(1) \equiv \widetilde{l}_1(1) = \frac{C + 1/\theta}{G^*(2\lambda)} + C_S. \tag{15}$$

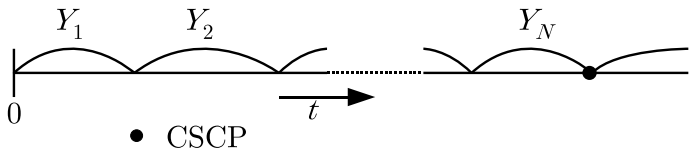


Fig. 3 Task execution for Scheme 2

(2) Scheme 2

CSCP is placed only at the end of task N ($N = 1, 2, \ldots$) in Fig. 3: When two states of all task k ($k = 1, 2, \ldots, N$) match at the end of task N, its state is stored and two modules execute task $N + 1$. When two states do not match, two modules go back to the first task 1 and make their retries. By the method similar to obtaining (15), the mean execution time of the process of all task N is

$$\widetilde{l}_2(N) = NC + \int_0^\infty \left\{ e^{-2\lambda t} (C_S + t) + \left(1 - e^{-2\lambda t}\right) \left[t + \widetilde{l}_2(N) \right] \right\} dG^{(N)}(t), \quad (16)$$

where $G^{(N)}(t)$ is the N-fold Stieltjes convolution of $G(t)$ with itself, i.e.,

$$G^{(N)}(t) \equiv \int_0^t G^{(N-1)}(t - u)\, dG(u) \qquad (N = 1, 2, \ldots),$$

and $G^{(0)}(t) \equiv 1$ for $t \geq 0$. Solving the above renewal equation for $\widetilde{l}_2(N)$,

$$\widetilde{l}_2(N) = \frac{N(C + 1/\theta)}{[G^*(2\lambda)]^N} + C_S.$$

Therefore, the mean execution time per one task is

$$l_2(N) \equiv \frac{\widetilde{l}_2(N)}{N} = \frac{C + 1/\theta}{[G^*(2\lambda)]^N} + \frac{C_S}{N} \qquad (N = 1, 2, \ldots), \quad (17)$$

which agrees with (15) when $N = 1$.

We find optimal number N_2^* to minimize $l_2(N)$ in (17). Forming the inequality $l_2(N + 1) - l_2(N) \geq 0$,

$$\frac{N(N + 1)[1 - G^*(2\lambda)]}{[G^*(2\lambda)]^{N+1}} \geq \frac{C_S}{C + 1/\theta} \qquad (N = 1, 2, \ldots), \quad (18)$$

whose left-hand side increases strictly with N to ∞. Thus, there exists a finite and unique minimum N_2^* ($1 \leq N_2^* < \infty$) which satisfies (18). If

$$\frac{2[1 - G^*(2\lambda)]}{[G^*(2\lambda)]^2} \geq \frac{C_S}{C + 1/\theta},$$

then $N_2^* = 1$. When $G(t) = 1 - \mathrm{e}^{-\theta t}$, (18) is

$$N\,(N+1)\,\frac{2\lambda}{\theta}\left(\frac{2\lambda}{\theta}+1\right)^N \geq \frac{C_S}{C+1/\theta}. \tag{19}$$

Example 4 Table 4 presents optimal N_2^*, the mean execution times $\theta l_2(N_2^*)$ and $\theta l_1(1)$ in (15) for λ/θ and θC when $\theta C_S = 0.1$. This indicates that N_2^* decrease with λ/θ and increase with θC. For example, when $\lambda/\theta = 0.005$ and $\theta C = 0.01$, $N_2^* = 3$ and $\theta l_2(N_2^*)$ is 1.077 which is about 4% smaller than $\theta l_1(1) = 1.120$ for Scheme 1. □

(3) Scheme 3

CSCP is placed at the end of task N and CCP is placed at the end of task k ($k = 1, 2, \ldots, N-1$) between CSCPs in Fig. 4: When two states of task k ($k = 1, 2, \ldots, N-1$) match at the end of task k, two modules execute task $k+1$. When two states of task k ($k = 1, 2, \ldots, N$) do not match, two modules go back to the first task 1. When two states of all task N match, the process is completed, and its state is stored. Two modules execute task $N+1$.

Let $\widetilde{l}_3(k)$ be the mean execution time from task k to the completion of task N. Using the similar method in Scheme 2,

Table 4 Optimal N_2^*, mean execution time $\theta l_2(N_2^*)$ and $\theta l_1(1)$ when $\theta C_S = 0.1$

$\frac{\lambda}{\theta}$	$\theta C = 0.005$			$\theta C = 0.01$			$\theta C = 0.05$		
	N_2^*	$\theta l_2\,(N_2^*)$	$\theta l_1\,(1)$	N_2^*	$\theta l_2\,(N_2^*)$	$\theta l_1\,(1)$	N_2^*	$\theta l_2\,(N_2^*)$	$\theta l_1\,(1)$
0.1	1	1.306	1.306	1	1.312	1.312	1	1.360	1.360
0.05	1	1.206	1.206	1	1.211	1.211	1	1.255	1.255
0.01	2	1.098	1.125	2	1.103	1.130	2	1.145	1.171
0.005	3	1.072	1.115	3	1.077	1.120	3	1.118	1.161
0.001	6	1.037	1.107	6	1.042	1.112	6	1.083	1.152
0.0005	8	1.028	1.106	8	1.033	1.111	7	1.074	1.151
0.0001	14	1.017	1.105	14	1.022	1.110	14	1.062	1.150

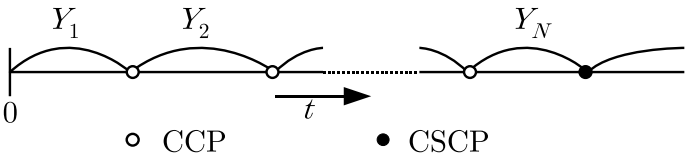


Fig. 4 Task execution for Scheme 3

$$\widetilde{l}_3(k) = C + \frac{1}{\theta} + \int_0^\infty \left[e^{-2\lambda t}\widetilde{l}_3(k+1) + \left(1 - e^{-2\lambda t}\right)\widetilde{l}_3(1)\right] dG(t)$$
$$(k = 1, 2, \ldots, N-1), \quad (20)$$

$$\widetilde{l}_3(N) = C + \frac{1}{\theta} + \int_0^\infty \left[e^{-2\lambda t} C_S + \left(1 - e^{-2\lambda t}\right)\widetilde{l}_3(1)\right] dG(t), \quad (21)$$

and solving them for $\widetilde{l}_3(1)$,

$$\widetilde{l}_3(1) = \frac{(C + 1/\theta)\left\{1 - [G^*(2\lambda)]^N\right\}}{[1 - G^*(2\lambda)][G^*(2\lambda)]^N} + C_S. \quad (22)$$

Therefore, the mean execution time per one task is

$$l_3(N) \equiv \frac{\widetilde{l}_3(1)}{N} = \frac{(C + 1/\theta)\left\{1 - [G^*(2\lambda)]^N\right\}}{N[1 - G^*(2\lambda)][G^*(2\lambda)]^N} + \frac{C_S}{N} \quad (N = 1, 2, \ldots), \quad (23)$$

which agrees with (15) when $N = 1$. Comparing (23) with $l_2(N)$ in (17), Scheme 3 is better than Scheme 2.

We find optimal N_3^* to minimize $l_3(N)$ in (23). Forming the inequality $l_3(N+1) - l_3(N) \geq 0$,

$$\frac{1}{[G^*(2\lambda)]^{N+1}} \sum_{k=1}^{N} \left\{1 - \left[G^*(2\lambda)\right]^k\right\} \geq \frac{C_S}{C + 1/\theta} \quad (N = 1, 2, \ldots), \quad (24)$$

whose left-hand side increases strictly with N to ∞. Thus, there exists a finite and unique minimum N_3^* $(1 \leq N_3^* < \infty)$ which satisfies (24). If

$$\frac{1 - G^*(2\lambda)}{[G^*(2\lambda)]^2} \geq \frac{C_S}{C + 1/\theta},$$

then $N_3^* = 1$. Comparing (24) with (18), it can be easily seen that $N_3^* \geq N_2^*$.

When $G(t) = 1 - e^{-\theta t}$, (24) is

$$\left(\frac{2\lambda + \theta}{\theta}\right)^{N+1} \sum_{k=1}^{N} \left[1 - \left(\frac{\theta}{2\lambda + \theta}\right)^k\right] \geq \frac{C_S}{C + 1/\theta}, \quad (25)$$

whose left-hand side increases with λ/θ from 0 to ∞. Thus, N_3^* decreases with λ/θ. If $(2\lambda/\theta)[(2\lambda/\theta) + 1] \geq C_S/(C + 1/\theta)$, then $N_3^* = 1$.

Example 5 Table 5 presents optimal N_3^* and the mean execution time $\theta l_3(N_3^*)$ in (23) for λ/θ and θC when $\theta C_S = 0.1$. For example, when $\lambda/\theta \geq 0.05$, $N_3^* = 1$ in all cases. It is clearly seen that Scheme 3 is better than Scheme 2 and $N_3^* \geq N_2^*$. However,

Table 5 Optimal N_3^* and mean execution time $\theta l_3(N_3^*)$ when $G(t) = 1 - \mathrm{e}^{-\theta t}$ and $\theta C_S = 0.1$

$\frac{\lambda}{\theta}$	$\theta C = 0.005$		$\theta C = 0.01$		$\theta C = 0.05$	
	N_3^*	$\theta l_3(N_3^*)$	N_3^*	$\theta l_3(N_3^*)$	N_3^*	$\theta l_3(N_3^*)$
0.1	1	1.306	1	1.312	1	1.360
0.05	1	1.206	1	1.211	1	1.255
0.01	3	1.085	3	1.090	3	1.132
0.005	4	1.060	4	1.066	4	1.107
0.001	7	1.031	7	1.036	7	1.076
0.0005	9	1.024	9	1.029	9	1.069
0.0001	17	1.015	17	1.020	17	1.060

overhead C for Scheme 2 would be less than that for Scheme 3. In such case, Scheme 2 might be better than Scheme 3. □

4.2 Majority Decision System

We take up a majority decision system with $(2n + 1)$ modules as an error masking system in (**3**) of Sect. 3 when $F(t) = 1 - \mathrm{e}^{-\lambda t}$. If more than $n + 1$ states of $(2n + 1)$ modules match, the process of task k is correct and its state is stored. In this case, the probability that the process is correct during $(0, t]$ is

$$\begin{aligned}\overline{F}_{n+1}(t) &= \sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \left(\mathrm{e}^{-\lambda t}\right)^j \left(1 - \mathrm{e}^{-\lambda t}\right)^{2n+1-j} \\ &= \sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \sum_{i=0}^{2n+1-j} \binom{2n+1-j}{i} (-1)^i \left(\mathrm{e}^{-\lambda t}\right)^{j+i} \\ &\qquad (n = 1, 2, \ldots). \end{aligned} \tag{26}$$

Thus, the mean execution time of the process of task k is

$$\widetilde{l}_4(1) = C + \frac{1}{\theta} + \int_0^\infty \left[\overline{F}_{n+1}(t) C_S + F_{n+1}(t) \widetilde{l}_4(1)\right] \mathrm{d}G(t), \tag{27}$$

and solving it for $\widetilde{l}_4(1)$, the mean execution time per one task for Scheme 1 is, from (14),

$$l_4(1) \equiv \widetilde{l}_4(1) = \frac{C + 1/\theta}{\sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \sum_{i=0}^{2n+1-j} \binom{2n+1-j}{i} (-1)^i G^*[(j+i)\lambda]} + C_S. \quad (28)$$

Similarly, the mean execution time per one task for Scheme 2 is, from (17),

$$l_5(N) = \frac{C + 1/\theta}{A(N)} + \frac{C_S}{N} \qquad (N = 1, 2, \ldots), \quad (29)$$

and the mean execution time per one task for Scheme 3 is, from (23),

$$l_6(N) = \frac{(C + 1/\theta)[1 - A(N)]}{N[1 - A(1)]A(N)} + \frac{C_S}{N} \qquad (N = 1, 2, \ldots), \quad (30)$$

where

$$\begin{aligned} A(N) &\equiv \int_0^\infty \overline{F}_{n+1}(t) \mathrm{d}G^{(N)}(t) \\ &= \sum_{j=n+1}^{2n+1} \binom{2n+1}{j} \sum_{i=0}^{2n+1-j} \binom{2n+1-j}{i} (-1)^i \left[G^*((j+i)\lambda)\right]^N. \end{aligned}$$

In particular when $n = 1$, i.e., for a 2-out-of-3 system, (28) is

$$l_4(1) = \frac{C + 1/\theta}{3G^*(2\lambda) - 2G^*(3\lambda)} + C_S, \quad (31)$$

and because

$$A(N) \equiv 3\left[G^*(2\lambda)\right]^N - 2\left[G^*(3\lambda)\right]^N,$$

$l_5(N)$ and $l_6(N)$ are easily given from (29) and (30), respectively.

Example 6 Suppose that $C \equiv \binom{2n+1}{2} C_1$ as same as Example 3. Tables 6 and 7 present optimal numbers N_5^* and N_6^* to minimize $l_5(N)$ and $l_6(N)$ in (29) and (30), respectively, and the mean execution times $\theta l_5(N_5^*)$ and $\theta l_6(N_6^*)$ for θC_1 and n, when $G(t) = 1 - \mathrm{e}^{-\theta t}$, $\lambda/\theta = 0.01$ and $\theta C_S = 1.0$. Table 6 indicates that N_5^* decrease with n, and $l_5(N_5^*)$ increase with n and θC. Tables 6 and 7 have the same tendencies. Noting that $l_5(N_5^*) < l_6(N_6^*)$, the optimal decision system is a 2-out-of-3 system. However, the overhead C might not increase with the order of n^2 of C_1. □

Table 6 Optimal N_5^* and mean execution time $\theta l_5(N_5^*)$ when $\lambda/\theta = 0.01$ and $\theta C_S = 1.0$

n	$\theta C_1 = 0.005$		$\theta C_1 = 0.001$		$\theta C_1 = 0.05$	
	N_5^*	$\theta l_5(N_5^*)$	N_5^*	$\theta l_5(N_5^*)$	N_5^*	$\theta l_5(N_5^*)$
1	47	0.058	40	0.079	25	0.223
2	36	0.100	30	0.158	19	0.585
3	31	0.159	26	0.272	17	1.137
4	28	0.236	24	0.423	17	1.885

Table 7 Optimal N_6^* and mean execution time $\theta l_6(N_6^*)$ when $\lambda/\theta = 0.01$ and $\theta C_S = 1.0$

n	$\theta C_1 = 0.005$		$\theta C_1 = 0.001$		$\theta C_1 = 0.05$	
	N_6^*	$\theta l_6(N_6^*)$	N_6^*	$\theta l_6(N_6^*)$	N_6^*	$\theta l_6(N_6^*)$
1	10	0.222	8	0.289	4	0.633
2	3	0.522	3	0.679	2	1.491
3	2	0.776	2	1.027	1	2.060
4	2	1.039	1	1.370	1	2.810

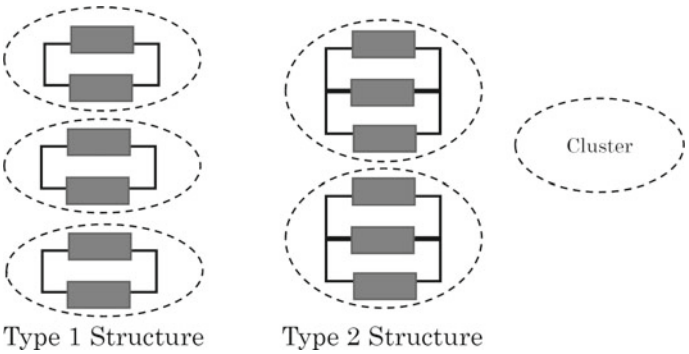


Fig. 5 Two types of structures

5 Random Checkpoint Models with Two Structures

When we have used a computer processing for multiple tasks, it would be impossible to improve its performance infinitely because we could not connect many modules. However, the performance could be improved greatly by preparing a lot of modules under the limitations of their number and power capacity. It is assumed that when there are 6 modules for Scheme 3 of Sect. 4, we adopt two types of structures which is a duplex system with 3 execution clusters and is a 2-out-of-3 majority system with 2 execution clusters in Fig. 5.

When modules have high-reliability, a duplex system can execute faster than a majority one. If there are a few errors in modules, then it needs a few rollbacks and

a duplex system can use more clusters than a majority one. On the other hand, when modules have low-reliability, a majority system can execute faster than a duplex one. If there are a lot of errors in modules, then a duplex system needs frequent rollbacks, but a majority system can mask a lot of such errors. Thus, it would be useful to obtain an intersection point of the mean execution times between two structures when each module has the same reliability.

(1) Type 1 Structure

We consider two checkpoint systems in which each of 3 clusters consists of 2 independent modules, executes and compares three states at checkpoint times. Each cluster adopts a duplex system with 2 modules as an error checking system in **(1)** of Sect. 3, and the system can execute 3 independent processing. Thus, the execution time is three times than that of one cluster. From (23), the mean execution time per cluster is

$$l_7(N) \equiv \frac{l_3(N)}{3} = \frac{(C+1/\theta)\left\{1-[G^*(2\lambda)]^N\right\}}{3N\,[1-G^*(2\lambda)]\,[G^*(2\lambda)]^N} + \frac{C_S}{3N} \quad (N=1,2,\ldots). \tag{32}$$

When $G(t) = 1 - \mathrm{e}^{-\theta t}$, (32) is

$$l_7(N) = \frac{(C+1/\theta)}{3N} \sum_{j=1}^{N} \left(\frac{2\lambda}{\theta}+1\right)^j + \frac{C_S}{3N}. \tag{33}$$

We find optimal number N_7^* to minimize $l_7(N)$. Forming the inequality $l_7(N+1) - l_7(N) \geq 0$,

$$\left(\frac{2\lambda}{\theta}+1\right)^{N+1} \sum_{j=1}^{N} \left[1-\left(\frac{1}{2\lambda/\theta+1}\right)^j\right] \geq \frac{\theta C_S}{1+\theta C} \quad (N=1,2,\ldots), \tag{34}$$

whose left-hand side increases strictly with N from $(2\lambda/\theta)\,[(2\lambda/\theta)+1]$ to ∞. Thus, there exists a finite and unique minimum $N_7^*(1 \leq N_7^* < \infty)$ which satisfies (34). Clearly, if

$$\frac{2\lambda}{\theta}\left(\frac{2\lambda}{\theta}+1\right) \geq \frac{\theta C_S}{1+\theta C},$$

then $N_7^* = 1$, i.e., we should provide no CCP between CSCP.

Example 7 Table 8 presents optimal number N_7^* in (34) and the mean execution time $\theta l_7(N_7^*)$ in (33) for λ/θ when $\theta C_S = 1.0$ and $\theta C = 0.5$. This indicates that N_7^* decrease with λ/θ to 1, and their mean execution times $l_7(N_7^*)$ increase with λ/θ. □

(2) Type 2 Structure

We consider two checkpoint systems in which each of 2 clusters consists of 3 independent modules, and executes and compares three states at checkpoint times [10].

Table 8 Optimal N_7^* and mean execution time $\theta l_7(N_7^*)$ when $\theta C_S = 1.0$, $\theta C = 0.5$

λ/θ	N_7^*	$\theta l_7\left(N_7^*\right)$
0.5	1	1.333
0.1	2	0.827
0.05	3	0.718
0.01	8	0.589
0.005	11	0.561
0.001	25	0.527
0.0005	36	0.519
0.0001	81	0.508

Each cluster adopts a majority decision system with 3 modules as an error masking system, i.e., 2-out-of-3 system in **(2)** of Sect. 3. Thus, the mean execution time is two times than that of one cluster. From (30), the mean execution time per cluster is

$$l_8(N) \equiv \frac{l_6(N)}{2} = \frac{(C+1/\theta)\left\{1-\left[3G^*(2\lambda)-2G^*(3\lambda)\right]^N\right\}}{2N[1-3G^*(2\lambda)+2G^*(3\lambda)]\,[3G^*(2\lambda)-2G^*(3\lambda)]^N} + \frac{C_S}{2N} \quad (N=1,2,\ldots). \tag{35}$$

When $G(t) = 1 - \mathrm{e}^{-\theta t}$, (35) is

$$l_8(N) = \frac{C+1/\theta}{2N}\sum_{j=1}^{N}\frac{1}{A^j} + \frac{C_S}{2N}, \tag{36}$$

where

$$A \equiv \frac{3\theta}{2\lambda+\theta} - \frac{2\theta}{3\lambda+\theta},$$

which decreases with λ/θ from 1 to 0.

We find optimal number N_8^* to minimize $l_8(N)$. Forming the inequality $l_8(N+1) - l_8(N) \geq 0$,

$$\frac{1}{A^{N+1}}\sum_{j=1}^{N}\left(1-A^j\right) \geq \frac{\theta C_S}{1+\theta C} \quad (N=1,2,\ldots), \tag{37}$$

whose left-hand side increases strictly with N from $(1-A)/A^2$ to ∞. Thus, there exists a finite and unique minimum N_8^* $(1 \leq N_8^* < \infty)$ which satisfies (37). If $(1-A)/A^2 \geq \theta C_S/(1+\theta C)$, then $N_8^* = 1$. It can be easily seen that N_8^* decrease with A, i.e., N_8^* decrease with λ/θ to 1.

Table 9 Optimal N_8^* and the mean execution time $\theta l_8(N_8^*)$ when $\theta C_S = 1.0$, $\theta C = 0.5$

λ/θ	N_8^*	$\theta l_8\left(N_8^*\right)$
0.5	2	1.551
0.1	5	0.945
0.05	10	0.851
0.01	48	0.771
0.005	95	0.761
0.001	472	0.752
0.0005	944	0.751
0.0001	4715	0.750

Example 8 Table 9 presents optimal number N_8^* in (37) and the mean execution time $\theta l_8(N_8^*)$ in (36) for λ/θ when $\theta C_S = 1.0$ and $\theta C = 0.5$. This shows a similar tendency to Table 8. Comparing Table 9 with Table 8, $N_8^* > N_7^*$ and $\theta l_8(N_8^*) > \theta l_7(N_7^*)$. □

5.1 Comparison of Two Structures

We compare Type 1 and Type 2 structures theoretically and numerically. Setting that $A \equiv \int_0^\infty \overline{F}(t)\mathrm{d}G(t) < 1$ in (22),

$$\frac{(C + 1/\theta)\left(1 - A^N\right)}{(1 - A)\,A^N} + C_S = \left(C + \frac{1}{\theta}\right)\sum_{j=1}^{N}\frac{1}{A^j} + C_S,$$

which decreases strictly with A from ∞ to $N(C + 1/\theta) + C_S$. When $F(t) = 1 - \mathrm{e}^{-\lambda t}$, because for $0 < t < \infty$,

$$3e^{-2\lambda t} - 2e^{-3\lambda t} > \mathrm{e}^{-2\lambda t},$$

we easily have

$$\int_0^\infty \left(3\mathrm{e}^{-2\lambda t} - 2\mathrm{e}^{-3\lambda t}\right)\mathrm{d}G(t) > \int_0^\infty \mathrm{e}^{-2\lambda t}\mathrm{d}G(t),$$

which follows that $l_3(N) > l_6(N)$. Thus, from (32) and (35), if $2l_3(N) < 3l_6(N)$, then $l_7(N) < l_8(N)$ and Type 1 structure is better than Type 2 one, and *vice versa*.

In particular, when $F(t) = 1 - \mathrm{e}^{-\lambda t}$ and $G(t) = 1 - \mathrm{e}^{-\theta t}$, if

Table 10 Values of λ/θ when $l_7(N) = l_8(N)$, $\theta C_S = 1.0$ and $\theta C = 0.5$

N	1	2	3	4	5	10	20	40	80	160	320
λ/θ	2.500	0.366	0.190	0.127	0.096	0.042	0.020	0.010	0.005	0.002	0.001

$$2\left(C+\frac{1}{\theta}\right)\sum_{j=1}^{N}\left(\frac{2\lambda}{\theta}+1\right)^{j}+2C_S$$
$$\leq 3\left(C+\frac{1}{\theta}\right)\sum_{j=1}^{N}\frac{1}{\{[3\theta/(2\lambda+\theta)]-[2\theta/(3\lambda+\theta)]\}^{j}}+3C_S,$$

i.e.,

$$\sum_{j=1}^{N}\left\{2\left(\frac{2\lambda}{\theta}+1\right)^{j}-3\left[\frac{(2\lambda+\theta)(3\lambda+\theta)}{5\lambda\theta+\theta^2}\right]^{j}\right\}\leq\frac{\theta C_S}{1+\theta C}, \tag{38}$$

then Type 1 structure is better than Type 2 one. For example, when $N = 1$, (38) is

$$\frac{2(\lambda/\theta)^2-\lambda/\theta-1}{5\lambda/\theta+1}\leq\frac{\theta C_S}{1+\theta C},$$

whose left-hand side increases with λ/θ from -1 to ∞. Letting x $(0 < x < \infty)$ be a solution of the equation

$$\frac{2x^2-x-1}{5x+1}=\frac{\theta C_S}{1+\theta C},$$

if $N = 1$ and $\lambda/\theta \leq x$, then Type 1 structure is better than Type 2 one.

Example 9 Table 10 presents the values of λ/θ for N when $l_7(N) = l_8(N)$, $\theta C_S = 1.0$ and $\theta C = 0.5$. This indicates that λ/θ decrease with N. When $N = 5$, if $\lambda/\theta < 0.096$, then Type 1 structure is better than Type 2 one, i.e., if error rates are small, we should adopt a duplex system, and so that, Type 1 structure is better than Type 2 one for the same N. □

Example 10 Figure 6 shows optimal numbers N_7^* and N_8^* of CCP checkpoints between CSCP for λ/θ when $\theta C_S = 1.0$ and $\theta C = 0.5$. This indicates that N_7^* and N_8^* decrease exponentially with λ/θ and N_7^* are smaller than N_8^*. □

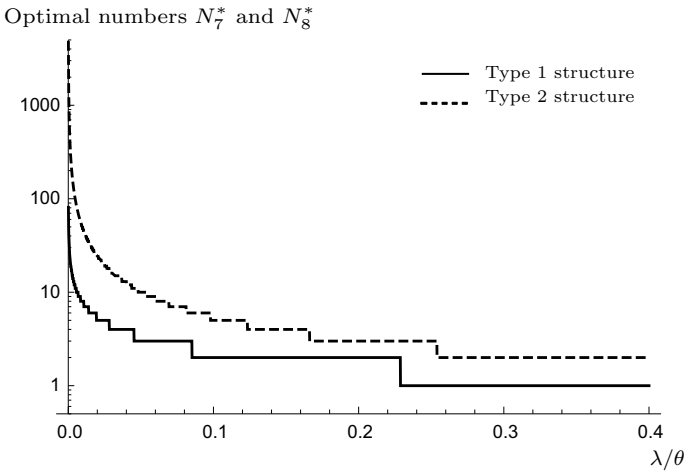


Fig. 6 Numbers N_7^* and N_8^* of CCP checkpoints between CSCP when $\theta C_S = 1.0$, $\theta C = 0.5$

6 Conclusion

We have studied optimal checkpoint intervals and structures with computing-modules for periodic and random checkpoint models. In particular, we have considered two types of structures in which three clusters consist of two modules and two clusters consist of three modules. It is of interest that if error rates are small, we should adopt a duplex system with three clusters. Furthermore, we have discussed the relationship between the error rates and the number of checkpoints, and have shown numerically that their optimal numbers are exponentially decreasing with error rates.

It has been assumed that the overheads for the generation of checkpoints and the probability of error occurrences are already known. However, it is important in practical fields to estimate statistically error rates and identify what type of distribution fits. If such error rates and distributions are given, we can determine optimal checkpoint policies for a real system by making some modifications and extensions of the methods and techniques used in this chapter. It would be necessary to formulate a computer system stochastically and to detect existing errors at appropriate checkpoint times.

This study would be applied to high-reliability requirement systems such as artificial satellite, medical system, auto-drive car, drone and so on, because these systems have to be provided with high-reliability by implementing redundant techniques and recovery methods.

References

1. Abd-El-Barr M (2007) Reliable and fault-tolerant. Imperial Collage Press, London
2. Barlow RE, Proschan F (1965) Mathematical theory of reliability. Wiley, New York
3. Birolini A (1999) Reliability engineering theory and practice. Springer, New York
4. Kim H, Shin KG (1996) Design and analysis of an optimal instruction-retry policy for TMR controller computers. IEEE Trans Comput 45(11):1217–1225
5. Kuo W, Prasad VR, Tillman FA, Hwang CL (2001) Optimal reliability design. Cambridge University Press, Cambridge
6. Lala PK (1985) Fault tolerant and faults testable hardware. Prentice-Hall, London
7. Lala PK (2001) Self-checking and fault tolerant digital design. Morgan Kaufmann Pub, San Francisco
8. Lee PA, Anderson T (1990) Fault tolerance principles and practice. Dependable computing and fault-tolerant systems. Springer, Wien
9. Levitin G (2007) Computational intelligence in reliability engineering. Springer, Berlin
10. Nakagawa T (2008) Advanced reliability models and maintenance policies. Springer, London
11. Nakagawa T, Yasui K, Sandoh H (2004) Note on optimal partition problems in reliability models. J Qual Maint Eng 10:282–287
12. Nakamura S, Nakagawa T (2010) Optimal checkpoint intervals for computer systems. World Scientific
13. Nanya T (1991) Fault tolerant computer. Ohm Co., Tokyo
14. Naruse K, Maeji S (2014) Chapter 13: Optimal checkpoint times for database systems. World Scientific, pp 231–247
15. Naruse K, Maeji S (2017) Optimal checkpoint times for redundant nodes and errors. World Scientific, pp 363–377
16. Naruse K, Nakagawa T (2019) Optimal structure of computing-nodes with limited number. In: 25th ISSAT international conference on reliability and quality in design, pp 53–57
17. O'Connor P (2001) Test engineering. Wiley, Chichester
18. Ohara M, Suzuki R, Arai M, Fukumoto S, Iwasaki K (2006) Analytical model on hybrid state saving with a limited number of checkpoints and bound rollbacks(reliability, maintainability and safety analysis). IEICE Trans Fundam Electron Commun Comput Sci 89(9):2386–2395. http://ci.nii.ac.jp/naid/110007537954/
19. Osaki S (1992) Applied stochastic system modeling. Springer, Berlin
20. Pradhan D, Vaidya N (1992) Roll-forward checkpointing scheme: concurrent retry with non-dedicated spares. In: Proceedings of the IEEE workshop on fault-tolerant parallel and distributed systems, pp 166–174
21. Siewiorek DP, Swarz RS (eds) (1982) The theory and practice of reliable system design. Digital Press, Bedford, MA
22. Sung C, Cho Y, Song S (2003) Combinatorial reliability optimization. Springer, pp 91–144
23. Ushakov IA (1994) Handbook of reliability engineering. Wiley, New York
24. Zhang Y, Chakrabarty K (2004) Dynamic adaptation for fault tolerance and power management in embedded real-time systems. ACM Trans Embed Comput 3(2):336–360
25. Ziv A, Bruck J (1997) Performance optimization of checkpointing schemes with task duplication. IEEE Trans Comput 46:1381–1386
26. Ziv A, Bruck J (1998) Analysis of checkpointing schemes with task duplication. IEEE Trans Comput 47:222–227

Kenichiro Naruse received B.S. degree from Aichi Institute of Technology in 1992, M.S. degree from Nagoya City University in 2003 and a Doctor degree from Aichi Institute of Technology in 2008. He has published 5 books chapters. He has interested in reliability of computer systems and IoT. He is working for Educational Affairs, Nagoya Sangyo University.

Toshio Nakagawa received B.S.E. and M.S. degrees from Nagoya Institute of Technology in 1965 and 1967, Japan, respectively; and a Ph.D. degree from Kyoto University in 1977, Japan. He worked as a Research Associate at Syracuse University for two years from 1972 to 1973. He is now a Honorary Professor with Aichi Institute of Technology, Japan. He has published 5 books from Springer, and about 200 journal papers. His research interests are optimization problems in operations research and management science, and analysis for stochastic and computer systems in reliability and maintenance theory.

Data Envelopment Analysis as a Tool to Evaluate Marketing Policy Reliability

Zaytsev Dmitry and Kuskova Valentina

Abstract In this paper we describe the Data Envelopment Analysis (DEA) research design and its applications for effectiveness evaluation of company marketing strategies. We argue that DEA is an efficient instrument for use in academia and industry to compare a company's business performance with its competitors'. This comparison provides the company with information on the closest competitors, including evaluating strategies with similar costs, but more efficient outcomes (sales). Furthermore, DEA provides suggestions on the optimal marketing mix to achieve superior performance.

Keywords Data envelopment analysis · Optimization · Business performance · Program evaluation

1 Introduction

Detailed descriptions of the method and its multiple extensions have been published in a variety of methodological references; a thorough breakdown is available, e.g., in Charnes et al. [6, 7] and Cooper et al. [10]. In brief, classical approach to DEA can be described as follows [7, 10, 14, 15]. There are K objects, or the decision-making units, each of which has multiple indices: a vector m of expended resources, or inputs, I_{km}, and the vector of n results or outputs, O_{kn}. Each resource has a weight of x_m, and each result—a weight of y_n in the formula for DEA efficiency calculation. It is important to note that the weight of each input or output is not known a priori. Using linear programming techniques, DEA algorithm solves a system of equations that optimizes a combination of weights, defining the object as either efficient or

Z. Dmitry (✉) · K. Valentina
International Laboratory for Applied Network Research, National Research University Higher School of Economics, Moscow, Russia
e-mail: dzaytsev@hse.ru

K. Valentina
e-mail: vkuskova@hse.ru

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_17

inefficient relative to other objects in the evaluation process. The efficiency of each object, formally, is the ratio of sum of all obtained results to the sum of all expended resources:

$$\frac{\sum_{n=1}^{N} O_{kn} y_n}{\sum_{m=1}^{M} O_{km} x_m} \tag{1}$$

Certain limits are imposed on the optimization algorithm. In basic DEA models, the value of the results cannot exceed the value of the expended resources. However, this assumption has been relaxed in more advanced models, called super-efficiency models (see, e.g., [16, 21]). A unit is 100% efficient if none of its inputs or outputs can be further improved without worsening other inputs or outputs. Adding new units into the mix may change the outcome of the estimation for any of the remaining units.

The resulting formulation of the optimization problem for the k-th object appears as follows. The first equation is the objective function, maximization of the efficiency of each evaluated unit, subject to imposed constraints (equations that follow).

$$\max z = \frac{\sum_{n=1}^{N} o_{kn} y_n}{\sum_{m=1}^{M} I_{km} x_m} \tag{2}$$

Subject to:

$$\sum_{n=1}^{N} O_{kn} y_n - \sum_{m=1}^{M} I_{km} x_m \leq 0 \quad k = 1, 2, \ldots, K \tag{3}$$

$$x_m, y_n \geq 0 \; for \; all \; m, n \tag{4}$$

Such objective functions are linear fractional, and they are linearized by making the denominator of the objective function equal to 1. The resulting optimization system looks as follows:

$$\max z = \sum_{n=1}^{N} O_{kn} y_n \tag{5}$$

Subject to:

$$\sum_{m=1}^{M} I_{km} x_m = 1 \tag{6}$$

$$\sum_{n=1}^{N} O_{kn} y_n - \sum_{m=1}^{M} I_{km} x_m \leq 0 \quad k = 1, 2, \ldots K \tag{7}$$

$$x_m, y_n \geq 0 \; for \; all \; m, n \tag{8}$$

Models can be output-oriented (as the system above, aimed at maximizing the output), or input-oriented (aimed at minimizing inputs). Input minimization is used when the purpose of research is to examine the extent to which resources can be reduced while maintaining the existing output levels. Output maximization looks at the levels to which the results can be raised given the current level of inputs. For the above system, the dual linear programming problem, minimizing inputs, is

$$\beta* = \min\beta_k \tag{9}$$

$$\sum_{k=1}^{K} O_{kn} z_k \geq O_{kn} \quad n = 1, 2, \ldots, N \tag{10}$$

$$\sum_{k=1}^{K} I_{km} z_k - I_{km} \beta_k \leq 0 \quad m = 1, 2, \ldots, M \tag{11}$$

$$z_k \geq 0 \, k = 1, 2, \ldots, K \tag{12}$$

The optimal value of β_k becomes the efficiency of the DMU k. By virtue of dual theorem of linear programming (LP), $z = \beta*$. This is derived from the foundational property of linear programming, where a dual of a LP is created from the original program. It is done in the following way: each variable in the primal LP is a constraint in a dual LP; each constraint in the primal LP is a variable in the dual LP; the objective direction is reversed—maximum in the primal is the minimum in the dual and vice versa [17]. Therefore, in DEA, either of the problems can be used, depending on the need, to evaluate efficiency as the ability to maximize outcomes or minimize inputs. As a result of the optimization procedure, we have a matrix of the optimum values of $\{x_{km}; y_{kn}\}$, and due to the imposed constraints, the resulting efficiency of each DMU cannot be greater than 1.

2 DEA Research Design

Variable selection and choice of return to scale are crucial for DEA research design. Both should be either theoretically driven or depend on an applied request from the client.

2.1 Variable Selection

In standard DEA, it is assumed that a variable constitutes either an input or an output. However, just as with regression methods, where selection of a dependent variable depends on a research context, in DEA some measures can play either input

or output roles. When theoretical roles of variables are not clear, a mathematical approach to variable selection may be necessary. Zhu and Cook [22] developed models that can help classify variables as either input or output; Cook et al. [8] provide the methodology for evaluating performance of the units where some factors can simultaneously play both the input and the output roles.

In addition, in real life, variables used for analysis are rarely homogeneous. For example, in the context of marketing research, such variables as internal company resources (choice of marketing mix, tangible and intangible resources) are discretionary for each company (meaning, the company has the ability to control their production and use), whereas the global competitive environment in which the company operates is non-discretionary. DEA allows for such heterogeneity in the data, providing an option for the researcher to select the non-discretionary indices that the DMU cannot control and appropriately exclude them from the calculation of an individual unit's efficiency, while allowing them to remain in the model for calculating the efficient frontier. Detailed description of models with non-discretionary inputs is provided, e.g., in Muñiz [18], Bessent et al. [4], and Banker and Morey [3].

2.2 Choice of Return to Scale

The economic concept of returns to scale (RTS) has further extended the applicability of DEA models [9]. Original extensions were suggested by Banker [1] and Banker et al. [2]. Nicholson [19] provides an intuitive explanation of constant returns to scale (CRS) as an "increase in outputs proportionate to the increase in inputs;" less than proportionate increase means diminishing returns, and more than proportional—increasing returns to scale (IRS). Most DEA software tools now provide options that incorporate all types of returns to scale for estimating the efficiency of the units; VRS models allow for units' returns to vary between the DMUs. Non-increasing returns to scale (NIRS) put an additional limitation on the VRS, where returns to scale for individual units are limited to constant or decreasing, forming a concave production function. While it is possible to determine the return to scale mathematically, there should be a theoretical foundation behind the model that determines that type of returns that one expects the DMUs to follow.

3 DEA-Generated Results

Depending on the program used, in addition to evaluation of efficiency, the output includes weights of each index in the efficiency calculations, slacks, improvements, graphical outputs, a subset of peer DMUs for each inefficient unit, and some others. Theoretical description of the most common outputs follows.

3.1 Efficiency

Efficiency is the relative performance indicator of a DMU in a group that it is being compared against. It is measured on a scale from 0 to 1, or as a percentage, depending on the software. Efficiency of one (100%) means that the unit is relatively efficient; anything less indicates that it is inefficient, with the actual value indicating the level of efficiency. Inefficiency means that either the inputs are not fully utilized, given the outcome, and could potentially generate a greater outcome (input-based models), or the outcome needs to be augmented (output-based models).

3.2 Weights

In the original model, Charnes et al. [7] accounts for the fact that DMUs may value inputs and outputs differently, so each unit is given the "freedom," mathematically, to vary the set of weights that allow it to achieve the maximum possible efficiency position relative to other units. The weights, therefore, are very flexible and unit-free, and are unrelated to any other input or output. If a unit is inefficient even with the most favorable choice of weights, relative to other units, then it is a strong statement to the unit's true lack of efficiency.

3.3 Slacks

Reducing inputs or augmenting outcomes alone may not always be sufficient. This situation is defined as a "slack," meaning excessive inputs or missing outputs. This result shows which resources are utilized completely to achieve a certain outcome, and which have an excess and could be reduced to improve performance.

3.4 Improvements (Targets)

One of the main benefits of DEA is estimating the magnitude of changes in the resource that is required for the unit to achieve the 100% efficiency. Improvements are provided in the form of changes to variables on the original scale. Some software packages may offer an option of a percentage change.

3.5 Peer Group (Peers, Reference Sets)

For every unit that is inefficient, DEA finds at least one unit that is efficient with the same basic pattern of inputs and outputs. Direct comparison of the peers is not always possible without additional data scaling, which can be done on either inputs or outputs. However, the peer group allows evaluating each inefficient unit's unique path to efficiency.

3.6 Cross-Efficiency

In addition to building frontiers to identify the most efficient cases, DEA constructed the cross-efficiency matrix—a table where the number of rows (i) and columns (j) equals the number of units in the analysis. For each cell (ij), the efficiency of unit ij is computed with weights that are optimal to unit j. The higher the values in a given column j, the more likely it is that the unit ij is an example of more efficiently operating practices. Doing so DEA allows us to match all units depending on the similarity of their efficiency function.

4 Evaluating Efficiency Over Time

DEA works with panel data. Because quite often scientists and practitioners are interested in the development of a phenomenon over time [13], this option can become invaluable for longitudinal efficiency comparison.

In classical productivity literature, changes in efficiency over time can be broadly divided into five categories [12]: (1) producing the same outputs with fewer resources, (2) producing more outputs without changing the resources used, (3) producing more outputs with fewer resources, (4) provide a larger increase in the outputs for an increase in inputs, and (5) provide a smaller reduction in outputs for the expected decrease in resources. The first three components are referred to as "technical efficiency," and the last two—as "scale efficiency." The total productivity growth achieved by the unit is estimated through the Malmquist index, first introduced by Caves et al. [5] and later extended by Fare et al. [11]. It is defined using the distance functions and does not require any assumptions regarding efficiency; therefore, it can distinguish between the factors causing chances in productivity. In a DEA setting, Malmquist index measures the change between data points by providing the ratio of distances from each data point to a common technology. As a result, it is decomposed into the technological change (TC) and efficiency change (EC). The latter is in turn is divided into the PEC (pure efficiency change) and SEC (scale efficiency change) indices for variable and constant returns to scale [20], respectively. Technical efficiency is the overall change in production technology, which reflects the movement

in the production frontier. Pure efficiency change component is measured relative to the true VRS frontier; where SEC is the "magnitude" component [11], which quantifies the productivity loss or gain of the unit itself. As a result, the Malmquist index is the product of change in relative efficiency between two different time periods, and the value of index equal to 1 implies no change in total factor productivity, less than 1—deterioration, more than 1—growth. Same interpretation applies to each individual component, and the different DMUs can be compared to each other directly in their changes of the Total Factor Productivity Change (TFPC).

5 An Application: Effectiveness of the Marketing Strategies

To illustrate the DEA capacities to evaluate efficiency of marketing strategy we choose an automotive company in Russia. The effectiveness of the company's marketing strategy is measured relative to other competing brands. In Table 1 we can see the efficiency scores, where effective brands get hundred, and non-efficient strategies have scores below hundred percent. We can also observe the change of efficiency in time. For example, marketing strategy of the car 9 was inefficient in 2017, and become efficient in 2018; and vice versa, brands of the cars 3 and 12 was inefficient in 2017 and become efficient in 2018. Finally, six brands have inefficient marketing strategies in both periods (Fig. 1).

Table 1 Marketing efficiency scores, %

Name	2017	2018
car 1	100	100
car 2	100	100
car 4	100	100
car 5	100	100
car 9	**69.05**	100
car 11	100	100
car 13	100	100
car 14	100	100
car 15	100	100
car 3	100	**83.58**
car 12	100	**76.13**
car 6	**81.3**	**70.6**
car 7	**77.9**	**69**
car 17	**56.98**	**57.63**
car 8	**85.78**	**57.19**
car 10	**68.25**	**53.29**
car 16	**11.79**	**21.69**

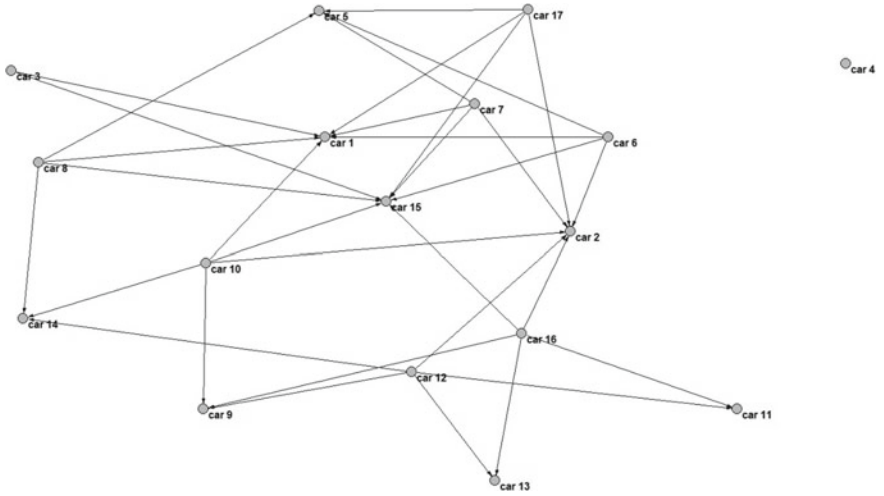


Fig. 1 Marketing peers, 2018

Tables 2 and 3 shows efficiency breakdown for the specific models of the analyzed brand with respect to spending via certain advertising channel. In Fig. 2, a one (1) against a channel shows that advertising through it was contributing to increased sales of an indicated model. To increase sales of the brands, they use advertising channels differently. For example, car 1 use regional TV channel only, and car 2 also use satellite TV. Regional TV is the most popular advertising channel to contribute in the cars sales. It is used by sixteen brands out of seventeen under study. Next popular advertising channels are outdoor marketing (10 cars use it) and satellite TV with 9 cars which use it.

In Table 3, the one against a channel for a specific model indicates that the company spent too much in resources for the level it achieved for a certain brand. When taken together, this information can help the company evaluate its marketing strategy. For example, spending on advertising in national TV was efficient for Cars 1, 9, 10, 12, 13–15; for all the other ten models, the company can cut national TV advertising without hurting the sales. DEA provides the exact "slack" numbers that could be cut without losing efficiency.

Table 4 presents marketing improvements or targets. For inefficient brands it shows the percentage of how much advertising spending should be reduced to increase sales. Also, it shows how sales will increase, in percentage, if the spending will be decreased respectively, as recommended. For example, car 3 have to cut digital advertising costs by approximately 19%, press advertising costs—by 72%, radio—by 19%, satellite TV—by 99%, national TV—by 40%, and regional TV—by 18%; which will lead to increase in sales by 20%.

Table 5 presents information for inefficient units about their peers. If brand is inefficient it means that it is under efficient frontier and the target position of this inefficient brand is a linear combination of the efficient brands. Therefore, each brand

Table 2 Marketing weights, 2018

Name	Digital	Outdoor	Press	Radio	Sat TV	TV Nat	TV Reg
car 1	0	0	0	0	0	0	1
car 2	0	0	0	0	1	0	1
car 3	0	1	0	0	0	0	0
car 4	0	0	1	0	0	0	1
car 5	0	1	0	0	0	0	1
car 6	0	1	0	0	0	0	1
car 7	0	1	0	1	0	0	1
car 8	0	0	1	0	1	0	1
car 9	0	1	0	0	0	0	1
car 10	0	0	0	0	1	1	1
car 11	0	1	0	0	1	0	1
car 12	0	0	0	0	1	1	1
car 13	1	1	0	0	1	0	1
car 14	0	0	1	0	1	1	1
car 15	0	1	0	0	1	1	1
car 16	1	1	0	0	1	0	1
car 17	0	1	0	0	0	0	1
Total	2	**10**	3	1	**9**	4	**16**

has two or more values (more than zero and less than one) related to its peers, which is called lambdas. Efficient brands have lambdas equal to one. For example, for car 3 we have car 1 and car 15 as it peers, which mean that the target position of the car 3 on the frontier is the linear combination of the spending and sales of car 1 and 15. If we will dichotomize lambdas in Table 5 in a way that zeroes remain zeroes and values more than zero will be ones, than we can get adjacency matrix for the network of marketing peers, Fig. 1.

Figure 1 shows a directed network, where the inefficient models are connected to the efficient—those which, with similar costs, achieved a greater sales result. Several things can be learned from this graph. First, connected car models are competitors to the efficient models; in other words, they use the same marketing strategy, but less efficiently. Second, because they follow the same marketing strategy, they are also competitors to each other—or, to similar inefficient models. Finally, this network can also be interpreted and analyzed as pseudo-bimodal, with one mode of "efficient" units, and another—of inefficient. Standard bimodal network analysis can then apply.

Table 6 present marketing cross efficiency values for each brand. We can interpret the values in the matrix in a way that each brand get efficiency scores calculated with the resources of related brand (e.g. car 3 with resources of car 1 will get 99.9% efficiency score; car 3 with car 2 resources – 20.4 efficiency score, etc.). If car has

Table 3 Marketing slacks, 2018

Name	Digital	Outdoor	Press	Radio	Sat TV	TV Nat	TV Reg
car 1	0	1	0	0	0	0	1
car 2	0	0	0	0	0	1	0
car 3	1	0	1	1	1	1	1
car 4	1	1	0	1	0	1	0
car 5	1	0	0	1	0	1	0
car 6	0	0	1	1	1	1	0
car 7	1	0	1	0	1	1	0
car 8	1	1	0	1	0	1	0
car 9	0	0	0	0	0	0	0
car 10	1	1	0	1	0	0	0
car 11	1	0	0	1	0	1	0
car 12	1	0	1	1	0	0	0
car 13	0	0	0	0	0	0	0
car 14	0	0	0	0	0	0	0
car 15	0	0	0	0	0	0	0
car 16	0	0	1	1	0	1	0
car 17	1	0	1	0	1	1	0
Total	**9**	**4**	**6**	**9**	**4**	**10**	**2**

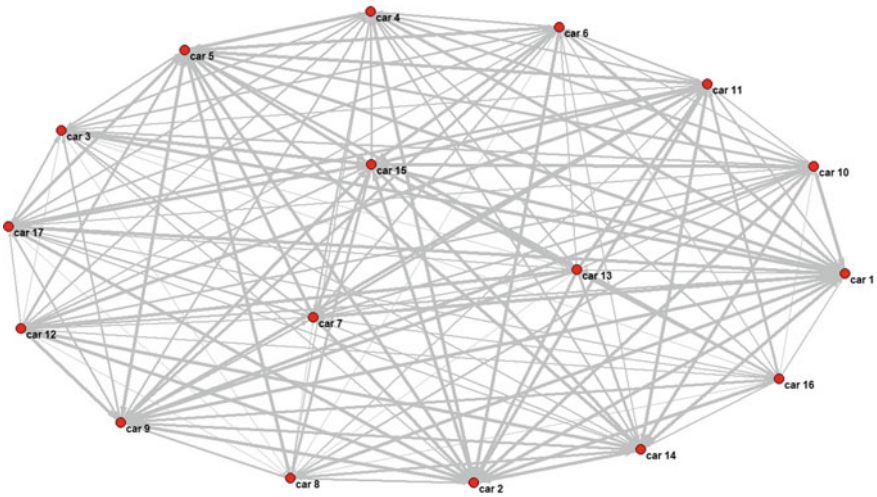


Fig. 2 Marketing cross efficiency, 2018

Table 4 Marketing improvements (targets), 2018, %

Name	Digital	Outdoor	Press	Radio	Sat TV	TV Nat	TV Reg	Sales
car 1	0	0	0	0	0	0	0	0
car 2	0	0	0	0	0	0	0	0
car 3	−18.76	0	−72.39	−19.03	−98.77	−39.99	−18.33	19.64
car 4	0	0	0	0	0	0	0	0
car 5	0	0	0	0	0	0	0	0
car 6	0	0	−43.31	−15.97	−76.12	−27.71	0	41.64
car 7	−39.11	0	−2.37	0	−71.57	−11.59	0	44.92
car 8	−86.88	−34.41	0	−86	0	−15.96	0	74.84
car 9	0	0	0	0	0	0	0	0
car 10	−63.82	−73.88	0	−58.95	0	0	0	87.64
car 11	0	0	0	0	0	0	0	0
car 12	−23.67	0	−22.25	−57.56	0	0	0	31.36
car 13	0	0	0	0	0	0	0	0
car 14	0	0	0	0	0	0	0	0
car 15	0	0	0	0	0	0	0	0
car 16	0	0	−78.14	−2.6	0	−63.91	0	361.09
car 17	−31.25	0	−66.7	0	−48.59	−23.94	0	73.52

efficient advertisement spending than its diagonal element is equal to 100. If not—below 100. The idea is that we get numbers that show similarities between brands not only for peers but for all brands. Let's take an example of comparison car 10 with car 1, here the score is more than hundred percent (100.1). So, the interpretation is that if car 10 will use resources of car 1 it will be more efficient compare to itself in 46.8%.

Table 6 is also adjacency matrix, so it can be presented as the weighted complete network, where we have weighed links from each car to each car, weights are the similarity values from Table 6. This network (Fig. 2) can be used to study further similarities between brands utilizing clustering and classification methods.

These indices show the increase/decrease of the efficiency of each unit between time periods. The change in efficiency of a brand other time can happened due to the change of efficiency of other brands (move of efficiency frontier); or due to the change of efficiency of a brand itself. TFPG index summarized both kinds of change. Therefore, efficient brands (have 100% efficient score stable through time) get only ones **in all indices** or get more than ones, mining that they even improve their position over time. Those brands where we observe decrease in efficiency (car 3, 6, 7, 8, 10, 12) have PEC index below one, and TC index more than one, mining that the decrease of their efficiency happened due to the worse performance relating to the brand itself and not to the frontier to which actually they become closer. Cars

Table 5 Marketing peers (lambdas), 2018

Name	car 1	car 2	car 4	car 5	car 9	car 11	car 13	car 14	car 15
car 1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
car 2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
car 3	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.41
car 4	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
car 5	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
car 6	0.10	0.18	0.00	0.34	0.00	0.00	0.00	0.00	0.38
car 7	0.06	0.16	0.00	0.56	0.00	0.00	0.00	0.00	0.22
car 8	0.17	0.00	0.00	0.03	0.00	0.00	0.00	0.23	0.56
car 9	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
car 10	0.07	0.12	0.00	0.00	0.04	0.00	0.00	0.26	0.51
car 11	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
car 12	0.00	0.03	0.00	0.00	0.18	0.12	0.38	0.30	0.00
car 13	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
car 14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
car 15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
car 16	0.00	0.00	0.00	0.00	0.04	0.07	0.03	0.00	0.87
car 17	0.17	0.46	0.00	0.18	0.00	0.00	0.00	0.00	0.20

Table 6 Marketing cross efficiency, 2018

Name	car 1	car 2	car 3	car 4	car 5	car 6	car 7	car 8	car 9	car 10	car 11	car 12	car 13	car 14	car 15	car 16	car 17
car 1	100.0	100.0	56.3	78.7	74.5	48.6	52.2	51.3	100.0	52.7	46.2	53.4	44.0	100.0	100.0	4.8	43.8
car 2	100.0	100.0	41.2	64.5	100.0	40.5	44.6	56.2	15.8	48.9	63.2	26.3	35.8	100.0	100.0	15.2	42.3
car 3	**99.9**	**20.4**	**83.6**	**11.3**	**55.2**	**26.9**	**26.5**	**29.3**	**17.6**	**4.5**	**26.7**	**6.4**	**2.5**	**10.8**	**95.0**	**10.1**	**17.6**
car 4	100.0	100.0	50.3	100.0	100.0	52.4	61.3	38.1	54.5	37.4	56.1	23.9	30.7	11.5	39.8	3.0	34.9
car 5	100.0	100.0	62.3	65.9	100.0	59.6	63.8	51.7	100.0	31.3	69.2	44.4	21.9	79.6	100.0	12.9	51.0
car 6	**99.8**	**99.7**	**69.6**	**61.6**	**99.3**	**70.6**	**56.5**	**27.1**	**87.2**	**21.9**	**69.9**	**32.5**	**17.7**	**37.8**	**90.6**	**27.0**	**52.6**
car 7	**99.9**	**99.8**	**69.9**	**56.2**	**99.7**	**67.8**	**69.0**	**28.3**	**60.8**	**20.8**	**76.5**	**18.3**	**11.7**	**31.5**	**94.3**	**20.1**	**57.7**
car 8	**98.9**	**54.6**	**20.3**	**28.8**	**96.3**	**18.9**	**20.9**	**57.2**	**4.3**	**29.2**	**43.2**	**9.2**	**15.0**	**76.4**	**71.8**	**11.7**	**24.0**
car 9	100.0	100.0	62.3	65.9	100.0	59.6	63.8	51.7	100.0	31.3	69.2	44.4	21.9	79.6	100.0	12.9	51.0
car 10	***100.1***	***100.2***	**56.4**	**78.9**	**74.8**	**48.8**	**52.4**	**51.8**	***101.1***	**53.3**	***46.9***	**54.7**	**45.8**	***105.0***	***108.4***	**6.5**	**44.0**
car 11	66.0	100.0	43.1	64.8	100.0	54.8	60.3	39.4	100.0	30.8	100.0	56.7	34.0	95.3	100.0	36.3	46.4
car 12	**41.2**	**99.6**	**26.8**	**68.3**	**75.8**	**40.3**	**45.7**	**25.1**	**97.8**	**31.9**	**96.9**	**76.1**	**93.6**	**92.6**	**50.9**	**20.9**	**35.8**
car 13	39.0	100.0	27.1	69.4	69.7	44.9	40.9	19.3	100.0	29.4	100.0	70.2	100.0	78.4	100.0	70.4	36.3
car 14	100.0	100.0	41.2	64.5	100.0	40.5	44.6	56.2	15.8	48.9	63.2	26.3	35.8	100.0	100.0	15.2	42.3
car 15	100.0	100.0	62.3	65.9	100.0	59.6	63.8	51.7	100.0	31.3	69.2	44.4	21.9	79.6	100.0	12.9	51.0
car 16	**37.9**	**95.5**	**25.5**	**64.3**	**63.3**	**39.8**	**36.2**	**15.4**	**78.0**	**22.4**	**71.5**	**44.8**	**54.2**	**39.5**	**42.8**	**21.7**	**33.0**
car 17	**99.9**	**99.8**	**69.9**	**56.2**	**99.5**	**67.7**	**68.9**	**28.3**	**60.6**	**20.7**	**76.1**	**18.1**	**11.6**	**31.0**	**92.1**	**19.1**	**57.6**

Table 7 Evaluating marketing efficiency over time (Ray and Desli Malmquist Index), 2017–2018

Name	TC	SEC	PEC	TFPG	First efficiency, 2017	Second efficiency, 2018
car 1	1	1	1	1	100	100
car 2	1	1	1	1	100	100
car 3	**1.16**	**1**	**0.84**	**0.97**	**100**	**83.58**
car 4	1.01	1.01	1	1.02	100	100
car 5	1	1	1	1	100	100
car 6	**1.42**	**1**	**0.87**	**1.23**	**81.3**	**70.6**
car 7	**1.44**	**1**	**0.89**	**1.28**	**77.9**	**69**
car 8	**1.52**	**0.97**	**0.67**	**0.98**	**85.78**	**57.19**
car 9	***1.14***	***1.46***	***1.45***	***2.42***	***69.05***	***100***
car 10	**1.47**	**1**	**0.78**	**1.14**	**68.25**	**53.29**
car 11	1	1	1	1	100	100
car 12	**1.15**	**0.97**	**0.76**	**0.85**	**100**	**76.13**
car 13	1	1	1	1	100	100
car 14	1	1.18	1	1.18	100	100
car 15	1	1	1	1	100	100
car 16	***2.71***	***0.38***	***1.84***	***1.88***	***11.79***	***21.69***
car 17	***1.44***	***1.02***	***1.01***	***1.48***	***56.98***	***57.63***

9, 16, and 17 increase their efficiency both based better performance related to itself and others (frontier) (Table 7).

6 Conclusion

Using the example of spending on marketing, we demonstrated the methodological capabilities of Data Envelopment Analysis, a non-parametric and mathematically rigorous method of evaluating effectiveness of different companies relative to each other, with a variety of inputs and outputs. DEA is rarely used in applied research compared to other popular methods such as linear regression. Yet, it has a number of advantages. First, as a non-parametric method, it is free from the limiting assumptions of data independence and identical distribution. Second, it can be used for data of all types, not just scaled on an interval or ratio levels, and can be applied to any combination of such data. Third, it can take a large number of both inputs and outputs. Because in applied research a large number of studies involve multiple independent and dependent variables, often measured on different scales and with different strengths of association, the usability of DEA should not be underestimated. This method overcomes many limitations that are inherent to data in the computer

science field, and we hope that researchers will take advantage of this tool to answer important questions not previously examined because of data issues.

Acknowledgements The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100.'

References

1. Banker RD (1984) Estimating most productive scale size using data envelopment analysis. Eur J Oper Res 17(1):35–44
2. Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 30(9):1078–1092
3. Banker RD, Morey RC (1986) The use of categorical variables in data envelopment analysis. Manag Sci 32(12):1613–1627
4. Bessent A, Bessent W, Kennington J, Reagan B (1982) An application of mathematical programming to assess productivity in the Houston independent school district. Manag Sci 28(12):1355–1367
5. Caves DW, Christensen LR, Erwin Diewert W (1982) The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica 50(6):1393–1414. https://doi.org/10.2307/1913388
6. Charnes A, Cooper WW, Lewin AY, Seiford LM (1994) Basic DEA models. In: Data envelopment analysis: theory, methodology, and applications. Springer, Dordrecht, Netherlands, pp 23–47
7. Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444
8. Cook WD, Green RH, Zhu J (2006) Dual-role factors in data envelopment analysis. IIE Trans 38(2):105–115
9. Cooper WW, Seiford LM, Zhu J (2004) Data envelopment analysis. Handbook on data envelopment analysis. Springer, Boston, MA, 1–39
10. Cooper WW, Seiford LM, Zhu J (eds) (2011) Handbook on data envelopment analysis, vol 164. Springer Science & Business Media
11. Färe R, Grosskopf S, Lindgren B, Roos P (1994) Productivity developments in Swedish hospitals: a Malmquist output index approach. In: Data envelopment analysis: theory, methodology, and applications. Springer, Dordrecht, Netherlands, pp 253–272
12. Golany B, Gang Y (1997) Estimating returns to scale in DEA. Eur J Oper Res 103(1):28–37
13. Goodin RE (2009) The state of the discipline, the discipline of the state. The Oxford handbook of political science
14. Khezrimotlagh D, Salleh S, Mohsenpour Z (2012) A new method in data envelopment analysis to find efficient decision making units and rank both technical efficient and inefficient DMUs together. Appl Math Sci 6(93):4609–4615
15. Khezrimotlagh D, Salleh S, Mohsenpour Z (2012) Benchmarking inefficient decision making units in DEA. J Basic Appl Sci Res 2(12):12056–12065
16. Lovell CAK, Rouse APB (2003) Equivalent standard DEA models to provide superefficiency scores. J Oper Res Soc 54(1):101–108
17. Matousek J, Gärtner B (2007) Understanding and using linear programming. Springer Science & Business Media
18. Muñiz MA (2002) Separating managerial inefficiency and external conditions in data envelopment analysis. Eur J Oper Res 143(3):625–643

19. Nicholson W (1985) Microeconomic theory: basic principles and extensions, 3rd edn. Dryden Press, Chicago, USA
20. Ray SC, Desli E (1997) Productivity growth, technical progress, and efficiency change in industrialized countries: comment. Am Econ Rev 87(5):1033–1039
21. Xue M, Harker PT (2002) Note: ranking DMUs with infeasible super-efficiency DEA models. Manag Sci 48(5):705–710
22. Zhu J, Cook WD (eds) (2007) Modeling data irregularities and structural complexities in data envelopment analysis. Springer Science & Business Media, Boston, MA, USA

Zaytsev Dmitry is an Associate Professor at the Faculty of Social Sciences, and a Senior Research Fellow at the International Laboratory for Applied Network Research of NRU Higher School of Economics, Moscow. He received his Ph.D. in political science from the Institute of World Economy and International Relations of the Russian Academy of Science. His current research interests cover the following topics: studies of democracy, social and political change; power, influence, and the impact of nation states; policy advisors and intellectuals as drivers of change; methodology in the political and social sciences. These topics incorporate current developments in political science, sociology, and social science methodology.

Kuskova Valentina is an Associate Professor at the Faculty of Social Sciences, Head of the International Laboratory for Applied Network Research, and Deputy First vice rector at NRU Higher School of Economics in Moscow, Russia. She received her Ph.D. in Management & Decision Sciences and a Master's in applied statistics from Indiana University. She is also the founder and Academic supervisor of the Master's Program "Applied Statistics with Network Analysis." Her research interests, in addition to research methods and network analytics, include methods of covariance structure analysis and non-parametric decision sciences.

Computational Intelligence Approaches for Software Quality Improvement

Grigore Albeanu, Henrik Madsen, and Florin Popenţiu-Vlădicescu

Abstract Obtaining reliable, secure and efficient software under optimal resource allocation is an important objective of software engineering science. This work investigates the usage of classical and recent development paradigms of computational intelligence (CI) to fulfill this objective. The main software engineering steps asking for CI tools are: product requirements analysis and precise software specification development, short time development by evolving automatic programming and pattern test generation, increasing dependability by specific design, minimizing software cost by predictive techniques, and optimal maintenance plans. The tasks solved by CI are related to classification, searching, optimization, and prediction. The following CI paradigms were found useful to help software engineers: fuzzy and intuitionistic fuzzy thinking over sets and numbers, nature inspired techniques for searching and optimization, bio inspired strategies for generating scenarios according to genetic algorithms, genetic programming, and immune algorithms. Neutrosophic computational models can help software management when working with imprecise data.

Keywords Computational intelligence · Immune algorithms · Software quality · Software reliability · Neutrosophic computational models

G. Albeanu
"Spiru Haret" University, Bucharest, Romania
e-mail: g.albeanu.mi@spiruharet.ro

H. Madsen
Danish Technical University, Lyngby, Denmark
e-mail: hmad@dtu.dk

F. Popenţiu-Vlădicescu (✉)
University Politehnica of Bucharest & Academy of Romanian Scientists, Bucharest, Romania
e-mail: Fl.Popentiu@city.ac.uk

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_18

1 Introduction

Improving quality of software is an important objective of any software developer. According to Jones [12], "software needs a careful analysis of economic factors and much better quality control than is normally accomplished".

The mentioned author has realized a deep analysis on software measurement and found out seven key metrics to be explored in order to estimate software economics and software quality with high precision [12]: the total number of function points, hours per function point for the project, software defect potential computed using function points, defect removal efficiency, delivered defects per function point; high-severity defects per function point; and security flaws per function point. For instance, the software defect potential is given by pow (number of function points, 1.25) [13], while the number of test cases can be estimated by pow (number of function points, 1.20), where pow (x, α) is x^{α}.

Many scientists, including Fenton and Pfleeger [9], Aguilar-Ruiz et al. [1], and Wójcicki and Dabrowski [38], have investigated software quality measurement using various metrics, statistical inference, and soft-computing methods.

Based on work [8] and the new developments in using artificial intelligence in software engineering [6, 22, 30, 33], this work considers classic and recent methods of Computational Intelligence (CI) applied in Software Engineering (SE) in order to identify improvements for solving the following tasks: software requirements analysis and precise software specification development; software development time reducing by evolving automatic programming and pattern test generation; dependability increasing by specific software design; and software cost/effort minimization by predictive techniques and optimal maintenance plans.

From the large variety of software quality definitions, in the following, the definition proposed by Jones [12] is used: "software quality is the absence of defects which would either cause the application to stop working, or cause it to produce incorrect results". As, software engineers proceed to develop a project, the main phases of the software life cycle are covered in this chapter.

The aim of this material, as an extension of [29], is to propose an extended approach based on fuzzy, intuitionistic fuzzy, and neutrosophic models for software requirement multi-expert evaluation (the third section), to update the usage of artificial immune algorithms for software testing (the fourth section), and to evaluate software reliability in neutrosophic frameworks (the fifth section).

2 Overview on Computational Intelligence Paradigms

According to IEEE Computational Intelligence Society [42], the main fields of Artificial Intelligence (AI) considered as special topics for CI are: Artificial Neural Networks (ANN), Fuzzy Systems (FS), Evolutionary Computation (EC), Cognitive and Developmental Systems (CDS), and Adaptive Dynamic Programming

and Reinforcement Learning (ADP&RL). Coverage of the main paradigms can be found in [16], while innovative CI algorithms are presented in [39]. The impact of computational intelligence on software engineering developments was revealed in [26].

Inspired by the biological network of neurons, Artificial Neural Networks are used as nonlinear models to classify data or to solve input-output relations [16]. Based on a weighted directed graph, having three types of vertices (neurons)—input, hidden and output, the ANN makes use of three functions for every vertex: network input f_{in}, neuron activation f_{act}, and output f_{out}. If the associated graph is acyclic then ANN is a *feed forward* network, otherwise is a *recurrent* network. The weights of the network are obtained by a training process.

One kind of learning task, called *fixed*, uses a set of pairs—*training patterns*—(x, y), where x is an input vector, and y is the output vector produced by ANN when received as input the vector x. Both x and y can be multivariate with different dimensions. The learning process is evaluated by some metric, like square root of deviations of actual results from desired output.

Another kind of learning task, called *free*, use only input vectors, the objective of ANN addressing a *clustering/classification* requirement. Here, a similarity measure is necessary to identify the prototypes. The power of ANN depends on the activation model of neurons, and the number of hidden layers. It is well known the results [16]: "any Riemann-integrable function can be approximated with arbitrary accuracy by a multilayer perceptron". Both practical and theoretical results on using different types of ANN have increased the confidence in using ANN as computational intelligence models.

For software engineering, the following references used ANN to optimize the software development process: Dawson [7] and Madsen et al. [18].

When the inputs are fuzzy [40, 41], intuitionistic fuzzy [3, 17], or of neutrosophic type [31], the activation process is based on defuzzification/deneutrofication procedures. Fuzzy systems make use of fuzzy sets, fuzzy numbers, and fuzzy logic. An intelligent FS is a knowledge based system used to answer to questions/queries formulated by a user according to a linguistic variables language. The natural language processing based interface is responsible on fuzzification/neutrofication procedure. Neutrosophic thinking for engineering applications is based on three indicators: one for *truth/membership* degree (T), one for the degree of *indeterminacy* (I), and one for *false/non-membership* degree (F). If $F + T = 1$, and $I = 0$, then the fuzzy framework [40] is considered. If $F + T < 1$, and $I = 1 - (F + T)$, then the intuitionistic-fuzzy theory of Atanassov is obtained [3]. The neutrosophic framework considers $0 \leq T + T + F \leq 3$, with $0 \leq T, I, F \leq 1$. Defuzzification can be obtained easy by the centroid method. However, many other methods were proposed and used in various contexts. Converting an intuitionistic-fuzzy entity, or a neutrosophic entity to a crisp value is not so easy. Firstly, an *indicator function* estimating the holistic degree of truth/membership should be computed (as in fuzzy representation), and this function will be used to compute a crisp value. The indicator function, denoted by H, is computed by

$$H = \alpha T + \beta(1 - F) + \gamma I/2 + \delta(1 - I/2),$$

for every item from universe of discourse. The parameters α, β, γ, δ are positive numbers, in decreasing order, with their sum being 1. The method was proposed by Wang et al. [36], and the parameters should be found by the researcher based on the available information about the problem under treatment.

Other neutrosophic computational models are presented and used in the fifth section.

FS systems use tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, low solution cost, and better rapport with reality, as Zadeh [42] recommended when he has coined the soft computing term.

Evolutionary computation is an area of research covering genetic algorithms, evolutionary strategies, and genetic programming. The techniques are based on a population of individuals and the following operations: *reproduction* (crossover/recombination), *random variation* (mutation, hypermutation), *competition*, and *selection*. The objective of any evolutionary algorithm is to optimize the searching process in a robust and intelligent manner, as inspired by biological reproduction schemes. Relevant results in software engineering were obtained, to mention some contributions, by Aguilar-Ruiz et al. [1], Arcuri and Yao [2], Berndt et al. [5], McGraw et al. [21], Michael et al. [23], Patton et al. [25], and Wappler and Lammermann [37].

Cognitive and Developmental Systems have specific targets in artificial life modelling and computational neuroscience: agent based modelling, special architectures of artificial neural networks, computational algorithms based on evolutionary strategies borrowed from real bio life. For software engineering, agent based modelling is one of CDS applications. The orientation on symbolic processing (as results of neural computing, concept transformation, and linguistic description) is the main reason to study CDS for software engineering. Also, the usage of machine learning for solving software engineering optimization problems motivates the application of CDS, as proposed by Wójcicki and Dabrowski [38] and Venkataiah et al. [35].

Adaptive Dynamic Programming and Reinforcement Learning are used to solve problems related to optimal control through efficient learning strategies. ADP&RL has to provide optimal decisions through knowledge based systems by active learning.

ADP&RL can be used in software engineering for optimal decision making along the software lifecycle, as described by Sultanov and Hayes [34].

3 Computational Intelligence for Software Requirements Engineering

The advancement in CI oriented technologies proved value in various SE specific applications: (1) automatic transformation of Natural Language Requirements into Software Specification; (2) software architecture design; (3) software coding and testing; (4) software reliability optimization; (5) software project management.

When consider Natural Language Requirements, the software engineer has to deal with ambiguity of requirements, incomplete, vague or imprecise descriptions, or the interpretation of the requirements. Requirements Ambiguity Reviews should be implemented at early phases of software development to obtain the following advantages [43]: requirements improvement and software defects reduction; 100% test coverage in order to identify software bugs; learn to differentiate between poor and testable requirements.

As Kamsties identified in [14], the requirements ambiguity reviewer should differentiate between *linguistic ambiguity* (context independent, lexical, syntactic) and *software engineering ambiguity* (context dependent, domain knowledge). If a sentence has more than one syntax tree (*syntactic ambiguity*), or it can be translated into more than one logic expression (*semantic ambiguity*) then an ambiguous requirement is found. When an anaphora or a pronoun has more than one antecedent, then a referential ambiguity should be processed. The analysts have also to do with pragmatic ambiguity generated by the relationship between the meaning of a sentence and its appearance place.

Special interest concerns the words like*: all, for each, each, every, any, many, both, few, some*, which are related to a whole set, or individuals in an unsized universe. Translating into logic expressions of sentences based on connectives like *and, or, if and only if, if then, unless, but* should address the truth membership degree and (intuitionistic) fuzzy norms, co-norms and implications. When detecting words like *after, only, with*, pronouns (*this, that, those, it, which, he, she*), *usually, often, generally, typically, normally*, and *also*, additional care is necessary when eliminates ambiguity. Other triggers announcing possibly ambiguity are given by *under-specified terms* (category, data, knowledge, area, thing, people etc.), *vague terms* (appropriate, as soon as possible, flexible, minimal, user-friendly, to the maximum extent, highly versatile etc.), and *domain-specific terms* (intelligence, security, level, input, source etc.). Osman and Zaharin described, in [24], an automated approach based on text classification to deal with ambiguous requirements.

However, automated disambiguation is impossible because human understanding is required to establish the requirements validity. In this case, a *multi-expert* approach is necessary to evaluate the requirements against ambiguity. The analysis proceeds in similar way for all fuzzy, intuitionistic-fuzzy, and neutrosophic approaches. A feasible strategy follows the steps:

1. Input one requirement as sentence in Natural Language or an Informal Language. Identify the requirement class (exact classification): Functional Requirements (FR), Nonfunctional Requirements (NFR), Performance/Reliability (PR), Interfaces (IO), Design constraints (DC).
2. After all requirements are considered, build the ambiguity degrees for every requirements class MFR[i], MNFR[i], MPR[i], MIO[i], and MDC[i] by every linguistic expert i, $i = 1, 2, \ldots, m$. The size of each matrix is given by the number of requirements identified for specified class. Consider a defuzzification/deneutrofication indicator, and for every requirement establish the tuple (r_k,

$type_k, e_{1,k}, e_{2,k}, \ldots, e_{m,k}$), where $e_{i,k}$ denotes the truth indicator function associated by expert i to requirement k.

3. Every requirement r_k, having $e_{i,k} \geq 0.5$ for at least one expert will be considered by software requirements engineer for lexical, syntactical, and semantical analysis in order to obtain a set of interpretations S_k. Contextualize every interpretation by a clear description.
4. Start the re-elicitation procedure against customer/client team, in order to establish the true software requirements.

One recent initiative in SRE is NaPIRE [10] which identifies the best practices used over the world. However, addressing the agile development paradigm, the practice already reported is not suited to other software development paradigms. The above proposal can be a setup for any paradigm, including the waterfall classic approach.

4 Computational Intelligence for Software Testing

There are a large variety of software testing methods and techniques which are different from the point of view of the paradigm type, effectiveness, ease of use, cost, and the need for automated tool support.

Evolutionary techniques can be applied to code development and for generating test cases, or unit testing. Also, evolutionary strategies apply when someone wants to estimate software development projects, as Aguilar-Ruiz et al. presented in [1].

The lifecycle of any evolutionary algorithm for software testing starts with a number of suitable test cases, as *initial population*. The evaluation metric is always based on a *fitness function*. Let be n the total number of domain regions/testing paths, k be the number of regions/paths covered by a test, hence the test case associated fitness/performance is k/n. The recombination and mutation operators applied to test case work as usual procedures applied to sequences/strings and are influenced only by the test case structure (representation).

Recently, Arcuri and Yao [2] use *co-evolutionary algorithms* for software development. In co-evolutionary algorithms, two or more populations co-evolve influencing each other, in a cooperative co-evolution manner (the populations work together to accomplish the same task), or in a competitive co-evolution approach (as predators and preys in nature).

Given a software specification S, the goal of software developer is to evolve a program P along some iterations in order to satisfy P. If genetic programming is used, the fitness of each program in a generation is evaluated on a finite set T of unit tests that depends on the specification S; the fitness value of a program being the maximum number of unit tests passed by the program. A better approach consists of using different sets T_i of unit tests for each new generation i.

The generation of new sets of unit tests can use the *negative selection* approach described in Popentiu and Albeanu [28]. If G_i is the set of genetic programs at step

i, and T_i is the set of unit tests for the i-th generation, the next generation is obtained according to the following method.

Let g(t) be the output of the program g having input t. Let c(t, g(t)) the fitness of the output of g related to the true output, when the precondition is valid, otherwise $c = 0$. If N(g) is the number of vertices of the program g (as flowchart), and $E_i(g)$ is the number of errors generated by g on t in T_i, the fitness degree of g is [2]:

$$f(g) = N(g)/(N(g) + 1) + E_i(g)/(E_i(g) + 1) + \Sigma\{c(t, g(t)); \text{ t in } T_i\}, \text{ g in } G_i.$$

Clonal selection can be used to test generation (a large collection of test cases can be obtained by mutation operator). The size of collection, considered like detectors, can be reduced by simulating a negative selection to eliminate those detectors which are not able to detect faults. The remaining detectors will be cloned and mutated, evaluated and used to create a new population of detectors.

The framework proposed by Arcuri and Yao can benefit from new algorithms for test case generation based both on genetic [5, 21, 23, 25] and immune [28] algorithms.

According to [20], for software testing and debugging, other models can be used:

1. Let K be the number of fault classes established by an expert, D be the program input domain, partitioned in n ($n > 1$) regions D_1, D_2, …, D_n, and Q be the probability distribution giving the operational profile: $\Sigma\{Q(x): x \in D\} = 1$. If $\phi_k(x)$ is the membership degree of x to the kth fault domain, and $P(f_k)$ is the probability to experience a kth type fault, then every software run will succeed with the degree R given by

$$R = 1 - \sum_{i=1}^{n} \int_{D_i} \left(\max_{j=1,\ldots,K} P\left(f_j\right)\phi_j(x)\right) Q(x)dx$$

   When the membership degrees φ_k are obtained by probability conversion, then R is the software reliability obtained in the probabilistic framework, according to Bastani and Pasquini [4]. Otherwise, R can be viewed as the membership degree when the universe of discourse contains all software items and the fuzzy set of the reliable items are considered.
2. Both the degree of detectability and the degree of risk can be estimated using fuzzy systems [20]. According to [11], the detectability of a test T is the probability that T is able to detect a bug in a selected software unit, if this software contains a bug. The degree of detectability of a testing approach T can be obtained by *mutation testing*. Also, the detectability can be given by a linguistic variable (*low, about low, average, about high, high*) with modifiers (hedges) like: *very, slightly, more-or-less* etc. In this way a situation like "the method T suspects a bug, more investigation are necessary" can be modelled and considered for a fuzzy rule database of an expert system for software testing.
3. Mutation is also a valuable operation used in the case of fuzz testing, or fuzzing [15]. According to [27], "fuzzing has long been established as a way to automate negative testing of software components". Mainly, fuzz testing is a technique for

software testing that generates random data to the inputs of a software unit in order to identify the existence of defects when the program fails.

5 A Neutrosophic Approach to Software Quality Evaluation

Neural networks and genetic algorithms can be used to provide an optimal reliability allocation in the case of modular design under fault tolerant constraints. These techniques were used by Madsen et al. [18, 19] in various contexts. In this section we use the neutrosophic numbers of Smarandache to evaluate the reliability/availability of software under a fault-tolerance design.

For our considerations, a *neutrosophic number* is an object of form $a + bI$, where a and b are real numbers, I is an operation such as $I^2 = I$, $I - I = 0$, $I + I = 2I$, $0I = 0$, with 1/I and I/I are not defined [31]. If $x = a + bI$ and $y = c + dI$ are neutrosophic numbers, then [32]:

(a) $x + y = (a + c) + (b + d)I$;
(b) $x - y = (a - c) + (b - d)I$;
(c) $xy = (ac) + (ad + bc + bd)I$;
(d) $\lambda x = (\lambda a) + (\lambda b)I$;
(e) $x/y = u + vI$ (when ever is possible), with $u = a/c$, and $v = (bc - ad)/(c^2 + cd)$.

In order to obtain u and v in (e), the following identification chain should be followed, according to the rules (a) and (c):

$$(u + vI)(c + dI) = (a + bI),$$
$$uc + (ud + cv + vd)I = (a + bI),$$
$$uc = a,\ u = a/c,\ ad/c + cv + vd = b,\ ad + c^2v + vcd = bc,$$
$$v\left(c^2 + cd\right) = bc - ad$$

The reliability of parallel structures is given by $R(A, B) = 1 - (1 - A)(1 - B) = X + IY$, with A, and B given by neutrosophic numbers, and $1 = 1 + 0I$. Therefore, when the reliability is appreciated by human experts as indeterminate, the new calculus permits the computation of the reliability, and the result may be interpreted as *minimum value* X, the *median* $X + Y/2$, the *first quartile* $X + Y/4$, or the *third quartile* $X + 3Y/4$ depending on the *deneutrofication procedure*.

The reliability of serial connected modules is given by $R(A, B) = AB$ (according to the rule c from above). This methodology can be used to evaluate the reliability of the bottom-up structures, and to transform reliability allocation/optimization problems formulated in neutrosophic manner. The solving strategy follows similar steps as describing by Madsen [17], and Madsen [19].

Neutrosophic numbers can be used for a neutrosophic estimation of the maintenance effort, based on a neutrosophic variant of the model of Belady and Lehman (cited by [9]):

$$M = p_1 + K(1 + d_1 - f_1) + I\big(p_2 + K(1 + d_2 - f_2)\big],$$

where $p_1 + Ip_2$ is the productive effort that involve analysis, design, coding, testing, and evaluation, $d_1 + Id_2$ is a complexity measure associated with poor design, $f_1 + If_2$ is the maintenance team unfamiliarity with the software, and K is an empirical constant. For instance, if the development effort was 480 + 20I persons per month, the project complexity was 7 + 2I, the degree of familiarity is 0.7 + 0.1I, with K = 0.25, then M = 481.825 + 24.75I which after deneutrofication by median can give the total effort expended in maintenance as 494.2 persons per month.

If a software will be reused after some code is rewritten, but the percentage of modified design (MD), the percentage of modified code (MC), the percentage of external code to be integrated (EI), amount of software understanding required (SU, computed taking into account the degree of unfamiliarity with software), the assessment and assimilation effort (AA) and the number of source lines of code to be adapted (ASLOC) are imprecisely known, and given by neutrosophic numbers, then applying the post-architecture model [43], it follows:

$$ESLOC = ASLOC\ (AA + SU + 0.4MD + 0.3MC + 0.3CI)/100.$$

Hence, the equivalent number of lines of new code (ESLOC) is obtained as a neutrosophic number. After deneutrofication, the crisp value can be obtained by *min* value or *quartile*—based scheme.

If ASLOC = 3200, AA = 2 + 2I, SU = 15 + 3I, DM = 15 + 5I, CM = 20 + 10I, CI = 50 + 20I, then ESLOC = 1408 + 512I.

In a similar way, other COCOMO equations [44] can be considered and used to derive various economical and quality indicators.

If the function points are used along the entire life cycle of software [13], and the number of function points between released versions are imprecisely known and modeled by neutrosophic numbers, the power of neutrosophic numbers is required.

Let be z = a + bI one neutrosophic number. To compute $z^{1.25}$, the following method can be used.

Since 1.25 = 5/4 it follows that $(a + bI)^{5/4} = (u + vI)$ and $(a + bI)^5 = (u + vI)^4$ for some u and v to be obtained.

Hence $a^5 + (5a^4b + 10a^3_b{}^2 + 10a^2b^3 + 5ab^4 + b^5)I = u^4 + (4u^3v + 6u^2v^2 + 4uv^3 + v^4)I$, with $u = a^{5/4}$, and $(a + b)^5 - a^5 = (u + v)^4 - u^4$. Therefore $(a + b)^5 = (u + v)^4$, with $v = (a + b)^{5/4} - a^{5/4}$.

In a similar way, we obtain $(a + bI)^{0.4} = a^{0.4} + ((a + b)^{0.4} - a^{0.4})$ I, a formula useful to derive the approximate development schedule in calendar months.

For a software having 2000 + 3000I function points, using the rules of Jones [13], it follows a development team of size 13.33 + 20I (about 23.33 full time personnel, using the median deneutrofication), while the size of maintenance team is 1.33 + 2I

(about 2.33 persons). The approximate development schedule in calendar months is 90.21 + 9.26I, about 25.54 months by median deneutrofication. Also, the software defect potential is given by $(2000 + 3000I)^{1.125} = 5172 + 9327I$, with a median deneutrofication value of about 9835.55.

Other rules formulated by Jones in [13], can be used in neutrosophic context.

6 Conclusions

This paper investigates the usage of classical and recent paradigms of computational intelligence to address some topics in software engineering: software requirements engineering, software design and software testing, software reliability allocation and optimization. Finally, neutrosophic computational schemes are proposed to evaluate various economic and quality indicators based on COCOMO model and Jones rules.

The proposed approach can be used with *min* deneutrofication operator and the results will be identical with those obtained by the classical approach (as an optimistic and theoretical view). However, using a *quartile* deneutrofication, the results can be closed to the values in practical software management.

References

1. Aguilar-Ruiz JS, Ramos I, Riquelme JC, Toro M (2001) An evolutionary approach to estimating software development projects. Inf Softw Technol 43(14):875–882
2. Arcuri A, Yao X (2014) Coevolutionary automatic programming for software development. Inf Sci 259:412–432
3. Atanassov KT (2016) Review and new results on intuitionistic fuzzy sets. Mathematical foundations of artificial intelligence seminar, Sofia, 1988, Preprint IM-MFAIS-1–88. Reprinted: Int J Bioautomation 20(S1):S7–S16
4. Bastani F, Pasquini A (1994) Assessment of a sampling method for measuring safety-critical software reliability. In: Proceedings of the 5th international symposium on software reliability engineering, 6–9 Nov, Monterey. IEEE Computer Society Press
5. Berndt D, Fisher J, Johnson L, Pinglikar J, Watkins A (2003) Breeding software test cases with genetic algorithms. In: Proceedings of the 36th annual Hawaii international conference on system sciences. https://doi.org/10.1109/HICSS.2003.1174917
6. Bisi M, Goyal NK (2015) Early prediction of software fault-prone module using artificial neural network. Int J Perform Eng 11(1):43–52
7. Dawson CW (1998) An artificial neural network approach to software testing effort estimation. Trans Inf Commun Technol 20. WIT Press
8. Feldt R, de Oliveira Neto FG, Torkar R (2018) Ways of applying artificial intelligence in software engineering. In: Proceedings of RAISE'18, Gothenburg, Sweden, ACM. https://doi.org/10.1145/3194104.3194109
9. Fenton NE, Pfleeger SL (1996) Software metrics: a rigorous and practical approach. Thomson
10. Fernandez DM (2018) Supporting requirements-engineering research that industry needs: the NaPiRE initiative. IEEE Softw 35(1):112–116
11. Howden WE, Huang Y (1994) Software trustability. In: Proceedings of the fifth International symposium on soft-ware reliability engineering, Monterey, California, pp 143–151. IEEE Computer Society Press

12. Jones TC (2017) The mess of software metrics. http://www.namcook.com/articles.html. Version 9.0
13. Jones TC (1998) Estimating software costs. McGraw-Hill
14. Kamsties E (2005) Understanding ambiguity in requirements engineering. In: Aurum A, Wohlin C (eds) Engineering and managing software requirements. Springer, Berlin, Heidelberg
15. Klees G, Ruef A, Cooper B, Wei S, Hicks M (2018) Evaluating fuzz testing. In: Proceedings of the ACM conference on computer and communications security, pp 2123–2138
16. Kruse R, Borgelt C, Braune C, Mostaghim S, Steinbrecher M (2016) Computational intelligence. A methodological introduction, 2nd edn. Springer
17. Madsen H, Albeanu G, Popenţiu Vlădicescu F (2012) An intuitionistic fuzzy methodology for component-based software reliability optimization. Int J Perform Eng 8(1):67–76
18. Madsen H, Thyregod P, Burtschy B, Albeanu G, Popentiu-Vladicescu F (2007) On using chained neural networks for software reliability prediction. In: Proceedings of the European safety and reliability conference 2007, ESREL 2007—risk, reliability and societal safety, vol 1, pp 411–417
19. Madsen H, Thyregod P, Burtschy B, Popentiu-Vladicescu F, Albeanu G (2006) On using soft computing techniques in software reliability engineering. Int J Reliab Qual Saf Eng 13(1):61–72
20. Madsen H, Thyregod P, Burtschy B, Albeanu G, Popentiu F (2006) A fuzzy logic approach to software testing and debugging. In: Guedes Soares C, Zio E (eds) Safety and reliability for managing risk, pp 1435–1442
21. McGraw G, Michael C, Schatz M (2001) Generating software test data by evolution. IEEE Trans Software Eng 27(12):1085–1110
22. Meziane F, Vadera S (2010) Artificial intelligence in software engineering—current developments and future prospects. In: Meziane F, Vadera S (eds) Artificial intelligence applications for improved software engineering development: new prospects, information science reference, pp 278–299. https://doi.org/10.4018/978-1-60566-758-4.ch014
23. Michael CC, McGraw GE, Schatz MA, Walton CC (1997) Genetic algorithms for dynamic test data generation. In: Proceedings 12th IEEE international conference automated software engineering. https://doi.org/10.1109/ASE.1997.632858
24. Osman MH, Zaharin MF (2018) Ambiguous software requirement specification detection: an automated approach, RET2018, June 2018, Gothenburg, Sweden http://ret.cs.lth.se/18/downloads/RET_2018_paper_9.pdf
25. Patton RM, Wu AS, Walton GH (2003) A genetic algorithm approach to focused software usage testing. In: Khoshgoftaar TM (ed) Software engineering with computational intelligence, vol 731. The Springer international series in engineering and computer science. Springer, Boston, MA, pp 259–286
26. Pedrycz W (2002) Computational intelligence as an emerging paradigm of software engineering. In: ACM proceedings of SEKE'02, pp 7–14
27. Petrică L, Vasilescu L, Ion A, Radu O (2015) IxFIZZ: integrated functional and fuzz testing framework based on Sulley and SPIN. Rom J Inf Sci Technol 18(1):54–68
28. Popentiu-Vladicescu F, Albeanu G (2017) Recent advances in artificial immune systems: models, algorithms, and applications. In: Patnaik S (ed) Recent developments in intelligent nature-inspired computing. IGI Global, Hershey, PA, pp 92–114. https://doi.org/10.4018/978-1-5225-2322-2.ch004
29. Popentiu-Vladicescu F, Albeanu G, Madsen H (2019) Improving software quality by new computational intelligence approaches. In: Pham H (ed) Proceedings of 25th ISSAT international conference "Reliability & Quality in Design", pp 152–156
30. Rech J, Althoff K-D (2004) Artificial intelligence and software engineering: status and future trends. Spec Issue Artif Intell Softw Eng 3:5–11
31. Smarandache F (2007) A unifying field in logics: neutrosophic logic. Neutrosophy, neutrosophic set, neutrosophic probability and statistics, 6th edn. ProQuest information & learning, Ann Arbor (2007)
32. Smarandache F (2016) Subtraction and Division of neutrosophic numbers. Crit Rev XIII:103–110

33. Sorte BW, Joshi PP, Jagtap V (2015) Use of artificial intelligence in software development life cycle—a state of the art review. IJAEGT 3(3):398–403
34. Sultanov H, Hayes JH (2013) Application of reinforcement learning to requirements engineering: requirements tracing. RE 2013, Rio de Janeiro, Brazil, Research Track. http://selab.netlab.uky.edu/homepage/publications/RE-hakim.pdf
35. Venkataiah V, Ramakanta M, Nagaratna M (2017) Review on intelligent and soft computing techniques to predict software cost estimation. IJAER 12(22):12665–12681
36. Wang H, Smarandache F, Zhang Y-Q, Sunderraman R (2005) Interval neutrosophic sets and logic: theory and applications in computing, Hexis
37. Wappler S, Lammermann F (2005) Using evolutionary algorithms for the unit testing of object-oriented software. In: GECCO '05 proceedings of the 7th annual conference on genetic and evolutionary computation. ACM https://doi.org/10.1145/1068009.1068187
38. Wójcicki B, Dabrowski R (2018) Applying machine learning to software fault prediction. e-Inform Softw Eng J 12(1):199–216. https://doi.org/10.5277/e-inf180108
39. Xing B, Gao W-J (2014) Innovative computational intelligence: a rough guide to 134 clever algorithms. Springer
40. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
41. Zadeh LA (1994) Fuzzy logic, neural networks, and soft computing. Commun ACM 37(3):77–84
42. Computational intelligence society page. https://cis.ieee.org/
43. How-harmful-can-be-ambiguous-software-requirements. Cigniti's requirement testing framework (RTF). https://www.cigniti.com/
44. COCOMO manual. https://sunset.usc.edu/research/COCOMOII/Docs/modelman.pdf

Grigore Albeanu received his Ph.D. in Mathematics in 1996 from University of Bucharest, Romania. During the interval 2004–2007 he was the head of the UNESCO IT Chair at University of Oradea, and participated as principal investigator in a NATO project. Grigore Albeanu has authored or co-authored over 120 papers and 10 educational textbooks in applied mathematics and computer science. Since 2007, he is professor of computer science at "Spiru Haret" University in Bucharest. His current research interests include different aspects of scientific computing, including soft computing, modelling and simulation, software reliability, virtual reality techniques and E-Learning.

Henrik Madsen received a Ph.D. in Statistics at the Technical University of Denmark in 1986. He was appointed Ass. Prof. in Statistics (1986), Assoc. Prof. (1989), and Professor in Mathematical Statistics with a special focus on Stochastic Dynamical Systems in 1999. In 2017 he was appointed Professor II at NTNU in Trondheim. His main research interest is related to analysis and modelling of stochastic dynamics systems. This includes signal processing, time series analysis, identification, estimation, grey-box modelling, prediction, optimization and control. The applications are mostly related to Energy Systems, Informatics, Environmental Systems, Bioinformatics, Biostatistics, Process Modelling and Finance. He has got several awards. Lately, in June 2016, he has been appointed Knight of the Order of Dannebrog by Her Majesty the Queen of Denmark, and he was appointed Doctor HC at Lund University in June 2017. He has authored or co-authored approximately 500 papers and 12 books. The most recent books are Time Series Analysis (2008); General and Generalized Linear Models (2011); Integrating Renewables in Electricity Markets (2013), and Statistics for Finance (2015).

Florin Popenţiu-Vlădicescu was born on 17 September 1950, graduated in Electronics and Telecommunications from University POLITEHNICA of Bucharest in 1974, holds a Ph.D. in Reliability since 1981. Also, he is Associated Professor to University "Politehnica" of Bucharest, Faculty of Automatic Control and Computer Science. He is the founder of the first "UNESCO

Chair of Information Engineering", in UK, established at City University London, in 1998. He published over 150 papers in international journals and conference proceedings. Also he is author of one book, and co-author of 4 books. He has worked for many years on problems associated with software reliability and has been Co-Director of two NATO Research Projects, in collaboration with Telecom ParisTech and Danish Technical University (DTU). He is an independent expert to the European Commission—H2020 programme, for Net Services—Software and Services, Cloud. He is currently Visiting Professor at renowned European and South Asian universities. He was elected Fellow of the Academy of Romanian Scientists in 2008.