Fausto Pedro García Márquez
Benjamin Lev *Editors*

Data Science and Digital Business

Springer

Data Science and Digital Business

Fausto Pedro García Márquez
Benjamin Lev
Editors

Data Science and Digital Business

Springer

Editors
Fausto Pedro García Márquez
ETSI Industriales de Ciudad Real
University of Castilla-La Mancha
Ciudad Real, Spain

Benjamin Lev
LeBow College of Business
Drexel University
Philadelphia, PA, USA

ISBN 978-3-319-95650-3 ISBN 978-3-319-95651-0 (eBook)
https://doi.org/10.1007/978-3-319-95651-0

Library of Congress Control Number: 2018947779

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

Contents

About the Editors

Fausto Pedro García Márquez obtained his European PhD with highest distinction. He has been awarded these prices: First International Business Ideas Competition 2017 Award (2017); Runner-up (2015), Advancement (2013) and Silver (2012) by the International Society of Management Science and Engineering Management (ICMSEM); Best Paper Award in the International Journal of Renewable Energy (Impact Factor 3.5) (2015). He works at UCLM as Full Professor (Accredited as Full Professor since 2013), Spain, Honorary Senior Research Fellow at Birmingham University, UK, Lecturer at the Postgraduate European Institute, and he has been Senior Manager in Accenture (2013-2014). He has published more than 150 papers (65% ISI, 30% JCR and 92% Internationals), some recognized as: "Renewable Energy" (as "Best Paper 2014"); "ICMSEM" (as "excellent"), "Int. J. of Automation and Computing" and "IMechE Part F: J. of Rail and Rapid Transit" (most downloaded), etc. He is the author and editor of 25 books (published by Elsevier, Springer, Pearson, Mc-GrawHill, Intech, IGI, Marcombo, AlfaOmega, and others) and he registered 5 patents. He is the Editor of 5 Int. Journals, Committee Member more than 40 Int. Conferences. He has been Principal Investigator in 4 European Projects, 5 National Projects, and more than 150 projects for Universities, Companies, etc. He is Director of www.ingeniumgroup.eu.

Benjamin Lev is the University Trustee Professor at LeBow College of Business, Drexel University. He holds a PhD in Operations Research from Case Western Reserve University. Prior to joining Drexel university Dr. Lev held academic and administrative positions at Temple University, University of Michigan-Dearborn and Worcester Polytechnic Institute. He is the Editor-in-Chief of OMEGA-The International journal of Management Science; Co-Editor-in-Chief of International Journal of Management Science and Engineering Management; and serves on several other journal editorial boards (Interfaces, IAOR, ORPJ, Financial Innovation, IDIM, IIE-Transactions, ERRJ, JOR,). He has currently faculty appointments at five Chinese Universities (Beijing Jiaotong University, Chengdu University, Nanjing University of Aeronautics and Astronautics, Nanjing University of Information Science and Technology, Nanjing Audit University). He has published ten books, numerous articles and organized many national and international conferences.

Introduction to Data Science and Digital Business

Fausto Pedro García Márquez and Benjamin Lev

This book combines the analytic principles of digital business and data science with business practice and big data. The interdisciplinary, contributed volume provides an interface between the main disciplines of engineering and technology and business administration. Written for managers, engineers and researchers who want to understand big data and develop new skills that are necessary in the digital business, it not only discusses the latest research, but also presents case studies demonstrating the successful application of data in the digital business.

Today, the world is getting smarter with the use of computing and mathematical methodologies. Many of the domains are now based on intelligent analysis and their interpretation as per the requirement of automation. For that purpose, many methodologies are in practice, including the field of data science. Data science is a multidisciplinary blend of data inference and algorithm designed to solve complex problems analytically. The demand and importance of an analytic has increased rapidly over the past few years 'Science of Analysis' is technically known as analytics; in other words, it is the analysis of information to provide timely valuable decisions. For organizations that have policies or intend to spread and enhance their business by means of data driven decision-making, data science is the secret ingredient. Projects based on data sciences can redeem more returns and benefits from development of data-based product as well as from providing guidance using data. Chapter "Data Management and Visualization Using Big Data Analytics" discusses the main concept of data management by using the big data analytics. It also discusses the methodologies used to manage the big data in different industries.

F. P. García Márquez (✉)
Ingenium Research Group, University of Castilla-La Mancha, Ciudad Real, Spain
e-mail: FaustoPedro.Garcia@uclm.es

B. Lev
LeBow College of Business, Drexel University, Philadelphia, PA 19104, USA
e-mail: bl355@drexel.edu

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_16

Chapter "Data Science and Digital Business" applies data science methods to analyze storm surge induced flood risks along the Mississippi Gulf Coast by presenting the spatial risk distribution of the study area, using the Geographic Information System (GIS) based visualization and quantifying the flood risk in statistical relationships with the risk related factors employing multiple linear regression analysis models. The data are retrieved and visualized for the residential blocks. The maximum surge elevation data are collected and validated against representative historical hurricane wind and storm surge data recorded by the Federal Emergency Management Agency and National Hurricane Center. The maximum surge height above the land surface is calculated based on the elevations and tide level in the Mississippi Gulf Coast Basin. The statistics models using the multiple regression analysis method characterize the significant relationships among these risk related variables. The direct loss coverage can be estimated using the models.

Recent advances in Artificial Intelligence, Semantic Web and intelligent interaction devices have made conversational interfaces increasingly popular. These advances in technologies including automatic speech recognition and synthesis, natural language understanding, and generation are result of decades of work in these areas to make possible a more natural and intuitive communication with machines.

Chapter "Data Science and Conversational Interfaces: A New Revolution in Digital Business" describes the potential of Data Science to improve the performance of conversational interfaces and increase the number of users of these interfaces. Following this cycle, the more people use these systems, more data is generated to learn their models and improve their performance, thus increasing the number of users and extending the possibilities for new applications in Digital Business.

Chapter "After 2017: Managers Exit and Banks Arise" describes how economic models are changed by blockchain infrastructures. Managements of corporations are debating whether to maximize profits or to minimize debt, and it represented in optimization models. Since blockchain technologies can automatically fulfill business contracts and achieve managements, managers will be treated as software. New economic agent, called "banks", is introduced in this chapter. The banks provide finance service and financial intermediation to other economic agents, and other agents provide electricity and computing resources to the banks. Also, the chapter points out that a new kind of corporations will arise. The corporations provide products whose marginal cost is zero by replacing the cost to accounting subject to other corporations and consumers.

Recommender system, famous for finding potential requirement of customers, is widely applied in many domain, such as bank, mobile, music, book and so on. Chapter "Synergies Between Association Rules and Collaborative Filtering in Recommender System: An Application to Auto Industry" proposes an integrated system containing data-processing, recommendation and evaluation processed. Traditional recommendation process aimed to recognize items that are more likely to be preferred by a specific user—however, it is expensive to recommend items to users who have no buying intention. Therefore, the chapter proposes a two-stage recommendation process by adopting advantages of many recommender technology. The first stage uses association rules as a means of classifying customers and finding potential

customers. In the second stage, collaborative filtering (CF) methods are applied to realize recommendation. In experimental study, the auto industry database is used to illustrate the proposed system. First, some implied information for the rules generated are set, which conforms to the observation. Second, CF method based on users' implicit preference information is developed.

Business digitization has aggravated the existing security and privacy concerns of customers, resulting in new challenges in organizational security and privacy protection. However, the knowledge on whether information security (ISec) studies respond in a timely manner to the requirements of industry and the situational changes in security and privacy brought about by digitizing business is limited.

Chapter "Information Security Research Challenges in the Process of Digitizing Business: A Review Based on the Information Security Model of IBM" compares the match between ISec papers published in six leading information system (IS) journals in the last four years and the themes of IBM ISec capability reference model. The chapter evaluates the practical relevance of ISec research in the IS field. Furthermore, four security objects (i.e., data, human behavior, IT/IS, and business processes) are identified in organizations and each of the six papers is coded to one or more of these security objects. By examining the interaction between security objects, it provides some suggestions for the research and industry communities.

Within the Data Science stack, the infrastructure layer supporting the distributed computing engine is a key part that plays an important role to obtain timely and accurate insights in a digital business. However, sometimes the expense of using such Data Science facilities in a commercial cloud infrastructure is not affordable to everyone. In this sense, Chapter "Deploying a Scalable Data Science Environment Using Docker" presents a computing environment based on free software tools over commodity computers. Thus, the chapter shows how to deploy an easily scalable Spark cluster using Docker including both Jupyter and RStudio that support Python and R programming languages. Moreover, it presents a successful case study where this computing framework has been used to analyze statistical results using data collected from meteorological stations located in the Canary Islands (Spain).

The world of marketing is becoming increasingly complex day by day with new channels and different ways of communication to interact with customers. The creativity in marketing and innovation in technology have pushed aside conventional ways of doing marketing. Customers are becoming more demanding. Right now, customers have the power and control to make their own decisions. There are three important things from the perspective of current marketing: creativity (uniqueness), strategy (game plan) and technology embedded with flexibility. Through these three contents marketers personalized consumer experiences can be achieved. Marketers should become aware of their customer's demands, and according to the business need, they adapt their e-CRM strategies to keep up with their concerned customers. In Pakistan, various new markets open with new vision, a business model which integrated with CRM technology optimizes the profitability and accessibility of physical and virtual stores in a country. Chapter "Clean up CHAOS and Use E-CRM (A Digital Concept) to Expand the Business: A Case of Pakistan" analyses the above issues.

In recent years, so-called "big data" has attracted the attention of companies and researchers. Chapter "Furthering Big Data Utilization in Tourism" aims to identify the number of visitors of each period and their characteristics based on the location data of mobile phone users collected by the mobile phone company. The study sites of this survey are tourist destinations in Ishikawa Prefecture and Toyama city, including Kanazawa city, which became nationally popular after the Hokuriku Shinkansen opened in 2015.

Generalized Additive logistic Regression model (GALRM) is a very important nonparametric regression model. It can be used for binary classification or for predicting the certainty of a binary outcome by using generalized additive models, which is known as a modern technique from statistical learning, and the penalized log-likelihood criterion. Chapter "C-LASSO Estimator for Generalized Additive Logistic Regression Based on B-Spline" develops an estimation problem for GALRM based on B-spline and Least Absolute Shrinkage and Selection Operator (LASSO). Unlike the traditional solutions, it will express the LASSO problem as a conic quadratic optimization problem, which is a well-structured convex optimization program, and solve it very efficiently using interior points methods.

A data-driven multi-objective bi-level signal design for urban network with hazmat transportation is considered in Chapter "An Efficient Bundle-Like Algorithm for Data-Driven Multi-objective Bi-level Signal Design for Traffic Networks with Hazardous Material Transportation". A bundle-like algorithm for a min-max model is presented to determine generalized travel cost for hazmat carriers under uncertain risk. A data-driven bi-level decision support system (DBSS) is developed for robust signal control under risk uncertainty. Since this problem is generally non-convex, a data-driven bounding strategy is developed to stabilize solutions and reduce relative gap between iterations. Numerical comparisons are made with other data-driven risk-averse models. The trade-offs between maximum risk exposure and travel costs are empirically investigated. As a result, the proposed model consistently exhibits highly considerable advantage on mitigation of public risk exposure whilst incurring less cost loss as compared to other data-driven risk models.

Analysis in data science and digital business requires analysis of the data, and in many cases the use of regression techniques. Chapter "Advanced Regression Models: Least Squares, Nonlinear, Poisson and Binary Logistics Regression Using R" discusses some simple regression and advanced regression techniques that have been used often in the analysis of data for business, industry, and government. Regression is not a one method fits all approach. Regression takes good approaches and common sense to complement the mathematical and statistical approaches used. It is also discussed methods to check for model adequacy after the regression model is found. The commands that were used in the examples are showed at the end of this chapter. It provides insights into the adequacy of the model through various approaches including regression ANOVA output, residual plots, and percent relative error. The chapter presents several methods to check for model adequacy.

Student attrition is a voluntary/involuntary failure or early dropout to complete a program in which an individual enrolled. For voluntary desertions, detection is more complex due to a variety of factors related to the program and individual context.

National academics have complained of a research shortage and desertion of student investigations in the Chilean context. Chapter "Student Desertion: What Is and How Can It Be Detected on Time?" applies data-mining techniques to reduce lack of studies and identify key factors and predict desertions during the first 6 semesters in a program of Business School at Universidad de Chile. 288 hybrid models were built, and the 6 final best models are composed of techniques of clustering, optimal-threshold classifiers, and SVM and Logistic Regression algorithms. In addition, the chapter shows the most important variables were related to University Selection Test (PSU in Spanish) score, followed by the educational level of parents and academic performances. The second level of importance included the variables of funding and family configuration.

In Japan, the aging society presents a significant problem. In 2014, a publication of the aging society published by the Japanese cabinet office, announced in October 2010 and October 2013 that the elderly constituted 23 and 25.1% of the population respectively. The average age is more than 65 years. This condition is the highest proportion in the world. In Chapter "Bioelectric Potential Plant", the condition of the elderly is mapped into two groups: the elderly who live with their families and who live alone. Based on data from samples taken in one of the major provinces in Japan, Kyoto, the number of the second group in 1990 was 43.416 (13.3%) and in 2010 the number increased to 110.366 (18.2%). These conditions lead to various problems one of which is that death of individuals often remains unknown by others, whether the death is caused by accidents in the home or other factors such as murder. Based on the same research, the deaths caused by accidents in the home without assistance was as much as 12.5%. This reality leads to the increasing demand for indoor monitoring. One of the measures being initiated is to examine the installation of closed circuit television (CCTV) cameras. This camera can monitor for accidents and ensure immediate help is obtained from a neighbor or an authorized officer. However, this solution is not well accepted due to privacy concerns. The use of infrared sensors was also tested to solve this problem. Despite the good results, costs are high because it requires many sensor cells. Then, other solutions have been tried by using the sense of odor, but the results are not effective because there is too much noise when the data records. Regarding this problem, Chapter "False Alarms Management by Data Science" proposes a solution through use of bioelectric potential sensor. This can be used for detecting human behavior and is friendly for use in private areas. In addition, the cost is reasonable. Therefore, this study outlines are explained by various methods such as machine learning and deep learning. Finally, the last chapter is the conclusion of discussion.

Due to the development of control system technology over the last years, the number of sensors has increased dramatically and the configuration of alarms in control systems has become easier. It leads to many alarms and increased operator workload. Industrial plants are currently underperforming due to alarm flood, which can cause minor, or even catastrophic, incidents. The businesses are demanding data science to avoid this, it is necessary to use process and alarm data. The industrial plants must understand the entire process and they rely on the experience of the operator. It has been considered that collaborative research between academic world

and industry should be undertaken to prevent flooding of alarms, both in normal and transitory conditions. New guidelines, standards and scientific/academic research should be developed. Nowadays new statistical, analytical and mathematical tools are being implemented for alarm detection, and the role of the operator must also be considered for correct alarm flood resolution. It will lead to a future with safer and more cost-effective industrial systems.

Data Management and Visualization Using Big Data Analytics

Muhammad Adeel Mannan, Saboohi Mehmood, Muhammad Shafiq and Aqeel-ur-Rehman

1 Introduction to Data Sciences

Nowadays, the realm is getting smarter with using computing and mathematical methodologies. Most of the domain names are now primarily based on smart evaluation and their interpretation as in keeping with the requirement routinely. For that purpose, among the methodologies are in practice that includes the sector of statistics sciences.

Analytics has risen quickly in popular business lingo over the last numerous years; the time period is used loosely, but usually intended to explain essential questioning that is quantitative in nature. Technically, analytics is the "science of analysis"—placed in other manner, the exercise of studying statistics to make decisions.

Now as per market scenario, most of the organizations and industries had revised their methodology and policies, and they rely on the changes which are statically driven, information technological knowledge is the name of the game sauce. As statically approach is more factual so any initiative will be more fruitful for both from steer through data insight, and development of statistics product.

From half-century ago, John Tukey nominated as a reformer of academic statistics. But in "The Future of Data Analysis" [1] he also pointed the data science as

M. Adeel Mannan · S. Mehmood · M. Shafiq · Aqeel-ur-Rehman (✉)
FEST, Department of Computing, HIET, Hamdard University, Karachi, Pakistan
e-mail: aqeel.rehman@hamdard.edu

M. Adeel Mannan
e-mail: adeel.mannan@hamdard.edu.pk

S. Mehmood
e-mail: saboohi.mehmood@hamdard.edu.pk

M. Shafiq
e-mail: m.shafiq@hamdard.edu.pk

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_1

unrecognized science although his area of interest was learning from data or data analysis. Approximately, twenty years ago, three scientists Bill Cleveland, Leo Breiman and John Chambers again works independently on academic statistics and expand their branches beyond the classical domains of theoretical statistics, Chambers nominated for preparation and presentation of data, Breiman called for predictions while Cleveland specified for data science.

So, according to Provost Marthan Pollack [1], "data science becomes the fourth approach in addition to experimentation modeling and computation".

Now a day's, context of information science has its multiple additives. Facts technological knowledge is movement itself and it inherently has to deal with the huge quantity of data. There are so many frameworks that are available that have their very own plus, minus but all of the most of the popular framework is Hadoop.

One part is very clear it's an extent massive quantity of the statistics which we are talking approximately. The second one part its miles the facts itself is huge verity; it is probably the information generated out of laws, out of sensors lets approximately airplane sensors, temperature sensors, it might be the records audio hills would be available.

The primary trouble is to keep the huge quantity of records and the second one is analysis method of data. This is one component. This is inventory change example and the opposite example speaks about jet flight. A half-hour flight generates 1 TB of statistics. The other thing is the speed we are accumulating and processing the records.

From time to time, the final deliverable is the kind of issue a statistician or commercial enterprise analyst may provide, however attaining that intention needs software competencies that an ordinary analyst surely doesn't have. As an instance, a dataset is probably so big that there is a need to apply allotted computing analysis on it or so convoluted in its layout that many strains of code are required to parse.

Here, two aspects of data science are of interest:

(i) The management and processing of information and
(ii) The analytical strategies and theories for descriptive and predictive evaluation and for prescriptive analysis and optimization.

The first issue entails statistics systems and their preparation, together with databases and warehousing, information cleaning and engineering, and facts tracking, reporting, and visualization. The second one issue includes facts analytics and consists of information mining, textual content analytics, device and statistical learning, possibility idea, mathematical optimization, and visualization [2].

This chapter discusses the primary concept of facts management by using the huge data analytics. It also discusses the methodology used to manage the massive records in extraordinary industries. Section 3 presents few actual international case studies. In segment four, some packages of facts sciences in the various area is beneath research even as in section five, we discuss some famous tools used for the control of huge records analytics. Later on, we have the conclusion of the chapter.

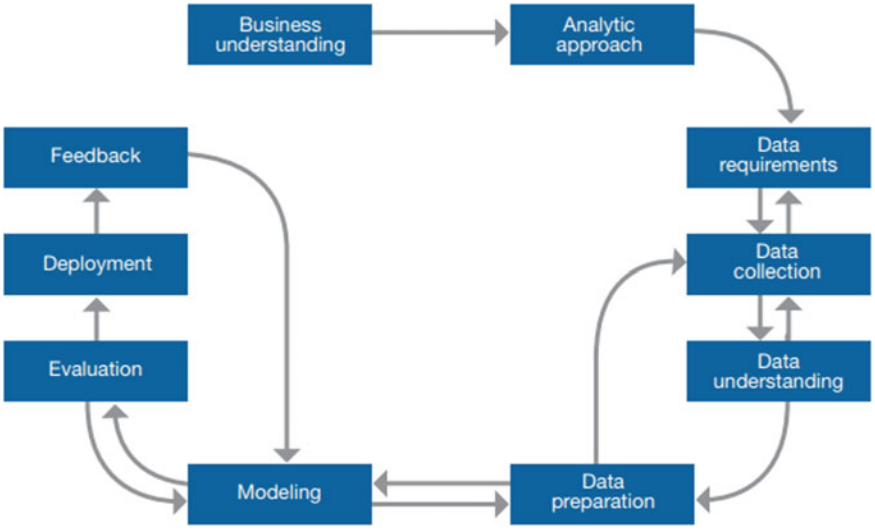


Fig. 1 Foundational methodology of data science

2 Data Management Methodologies

A general strategy (refer to Fig. 1) that guides the processes and actions with a given domain is called Methodology [3]. It does not base on particular tool/technology. Therefore, methodology gives the directions to data scientist within specific boundaries for how to proceed with the problem, what types of methods or strategies will be used that helps to reach on destination/results.

2.1 Business Understanding

The business level consists of placing objectives, generating assignment plans, and creating enterprise success standards. Those business sponsors who have analytical solutions play an important role at the time of the defining problem, its objectives, and solutions [3]. What is the pillar of the success of the business problem? Success for any project it is necessary for sponsors to take keen interest throughout the project and provide the best expertise [4].

2.2 Analytical Approach in Use

> "The use of analysis to break down the problem into its components for their solution is called analytical approach." [4]

The data scientist firstly states the problem clearly than define the analytical strategy to solving it. For expressing the problem in form of statistical and machine learning techniques it is compulsory the data scientist can identify the strategy that requires for achieving the suitable result.

2.2.1 Data Requirements

Data necessities are prescribed directives or consensual agreements that define the content material and/or structure that represent high fine statistics times and values. Data necessities can thereby be stated by using numerous exceptional individuals or corporations of people.

2.2.2 Data Collection

> "Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables want to answer relevant questions and evaluate outcomes." [5]

The data collection component of research is common to all fields of study. Huge data permits series of statistics with high scope and granularity. Because of advances in the era, statistics series techniques are more frequently limited by using the creativeness of the researcher than by using technological constraints [6]. In fact, one of the key demanding situations is to "assume out of doors the field" on the way to establish get entry to distinct facts for a huge variety of observations. Enormous statistics collection strategies that help to overcome this assignment encompass sensors, internet scraping, and internet visitors and communications tracking [7].

2.2.3 Data Understanding

The principal aim of information is to gain general insights approximately the information that will probably be beneficial for the similar steps within the data analysis technique; however, records knowledge must now not be pushed completely by using the goals and techniques to be applied in later steps. Despite the fact that those necessities have to be saved in mind at some point of information knowledge, one needs to approach the records from a neutral factor of view [8]. In no way trust any records as long as you have not performed some easy plausibility checks.

2.2.4 Modeling

Preliminary with the main type of the arranged data set, the demonstrating phase focuses on developing predictive or descriptive models in keeping with the formerly described analytic method [8]. With predictive models, records scientists use a schooling set (historic statistics wherein the outcome of the hobby is thought) to build the model. The modeling system is generally particularly iterative as corporations gain intermediate insights, leading to refinements in statistics guidance and version specification. For a given technique, information scientists can also endeavor a couple of procedures with their own constraints to locate the exceptional version for the variables.

2.2.5 Evaluation

The data scientist assesses the model to detain and make sure that it well and fully addresses the enterprise trouble [8]. Model assessment involves computing diverse diagnostic measures and different outputs which include tables and graphs, permitting the records scientist to interpret the model's great and its efficacy in solving the hassle. For a predictive model, information scientists use a tryout set, that's impartial of the training set, however, follows the identical chance distribution and has a regarded outcome. The testing set is used to evaluate the version so it can be difficult to detect as needed. Occasionally the final model has carried out additionally to a validation set for a very last assessment.

2.2.6 Deployment

The concept of deployment in facts science refers back to the utility of a model for prediction using some new statistics. Building a version is typically no longer the top of the task. Even if the reason for the model is to increase knowledge of the information, the knowledge received will need to be prepared and supplied in a manner that the consumer can use it. Relying on the requirements, the deployment phase may be as easy as producing a file or as complicated as enforcing a repeatable facts science process. In lots of cases, it is going to be the patron, not the statistics analyst, who will perform the deployment steps [9].

2.2.7 Feedback

Feedback to assess version overall performance;

- Collecting and analysis of feedback for evaluation of the model's overall performance
- An iterative procedure for version refinement and redeployment
- Boost up via automatic procedures.

2.3 Agile Kanban

Another methodology to work within data sciences is the integrated model of Agile along with the Kanban (a lean manufacturing domain for process management) [10]. It best suits while developing software based on data management. The key factor used here is the board named as "Kanban board", where we can track the progressive work easily.

This combined approach uses prioritized list of what to do? It means limit the task in terms of user stories. For example; in weather forecasting, comparison between new and previously formed data can limit the progressive work.

3 Real World Cases

Big data problems have diversified nature; let it be healthcare [11] or engineering problem, whether an insurance fraud discovery [12] or behavior analysis [13], whenever data analytics is involved, big data analytics techniques and data mining frameworks are used. In this section, Real world big data cases are covered in detail.

3.1 Case Study: Maritime Pattern Identification and Route Reconstruction

This study has discussed the problem of the arrival of deep-sea containers at Dutch terminals and proposed a solution to predict their arrival time with the publicly available AIS data [14]. This study aims at maximizing the utilization of inland water transportation resources (especially barges) and minimizing the waiting times for those barges within Dutch logistics service providers (LSPs). The occurrence of a terminal disturbance, a situation where the unscheduled arrivals of deep-sea containers at terminals make the barges wait, and terminals become unavailable for the barges. Hence predictions of these disturbances are favorable for efficient utilization of resources.

To predict the arrival time of a deep-sea container, this study makes use of the AIS data and with this data, sailing pattern extraction and transformation into a directed graph with a genetic algorithm is done. Thus, the main contribution of this study is a new approach on how to adopt a genetic algorithm to handle the real-life data mining problem of pattern extraction. It is important to know that why the genetic algorithm is used over other available techniques. First of all, data received in a sequential way from the sensors incrementally and real-time data may not always be accurate as the weather conditions are also affecting factor, therefore an algorithm which could give good results in the presence of all the above-mentioned issues was needed.

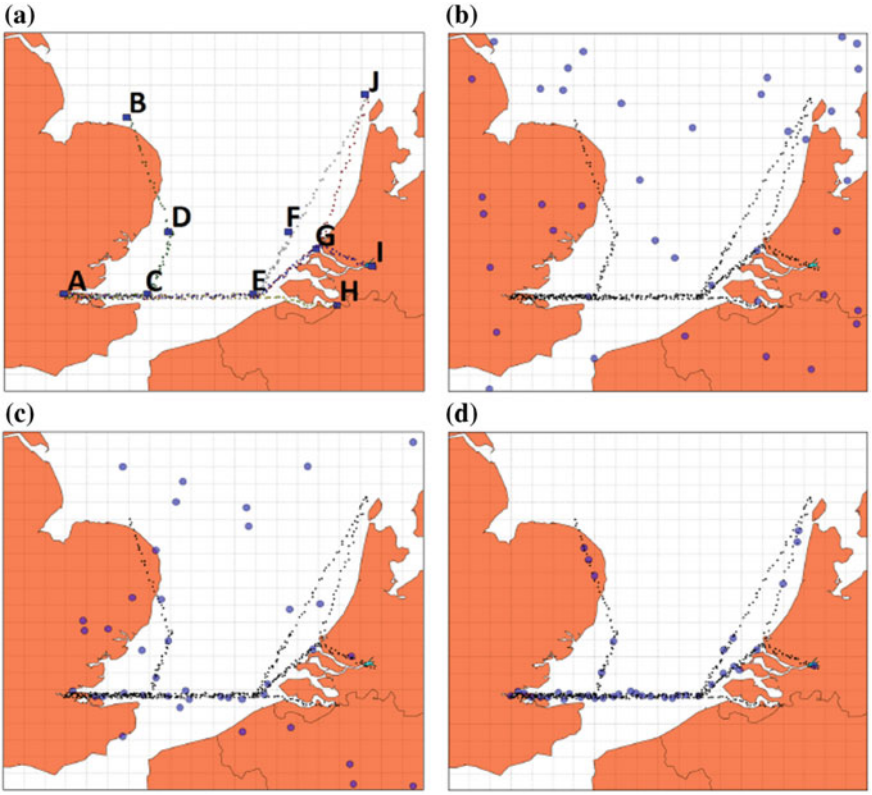


Fig. 2 Pictorial representation of the use of a genetic algorithm to identify routes of the vessels, arbitrary waypoints are converged towards a dotted area. [14]

The problem calls for finding a way to identify waypoints in maritime trajectories, i.e., special regions of interest where lanes intersect, as well as connecting them in such a way that it allows to depict all possible routes between the points. An illustration is shown in Fig. 2.

The steps needed to perform route extraction are depicted in Fig. 3. For route extraction, a directed graph is drawn which is based on the waypoints discovered by GA. Genetic algorithm finds the waypoints by first preprocessing available data with a quad tree; as quad tree structure has the ability to effectively identify good sets of clusters with varying data density, and also do it in only one pass. Once waypoints are discovered, a directed graph is generated after graph pruning and handling of missing data.

Once routes of deep-sea vessels are identified, their expected arrival time can be calculated, that is required for this case study. The results obtained from this study have shown that this approach is robust and lane identification even with the minimal

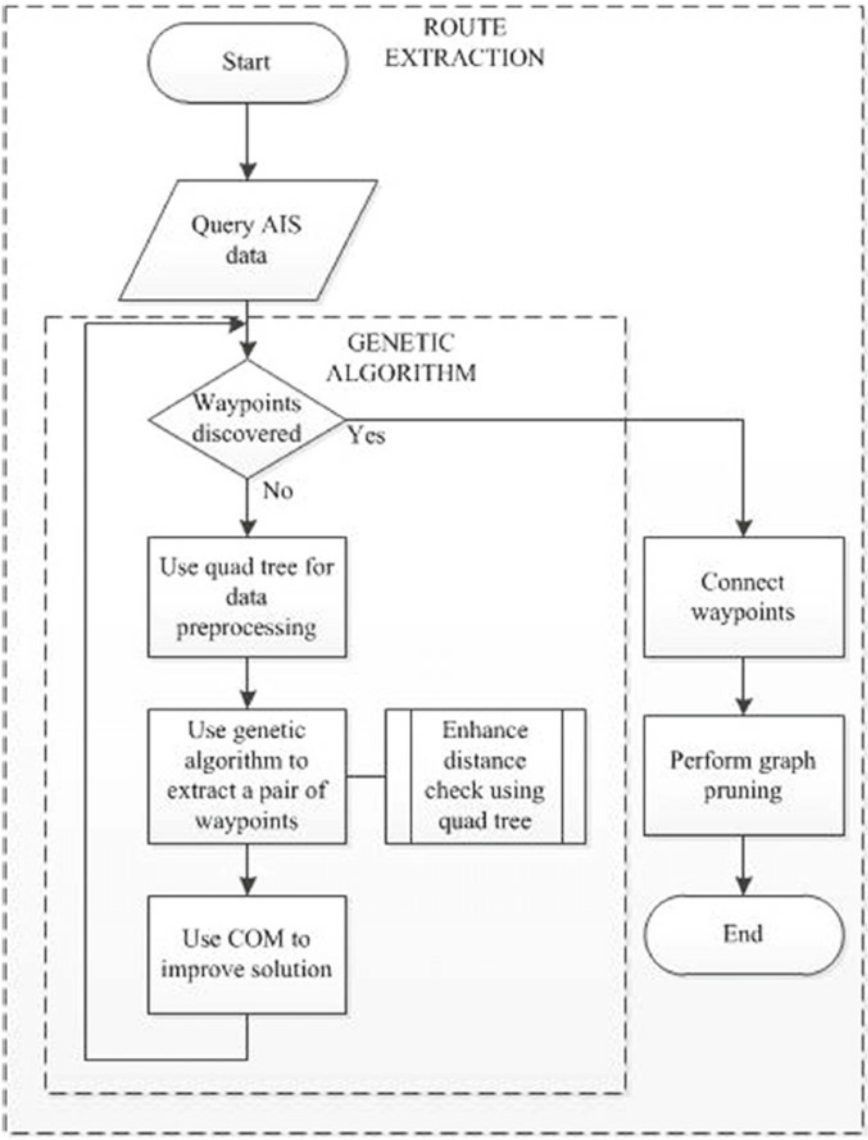


Fig. 3 Steps in improving pattern extraction [14]

data. Finally, the proposed solution is tested on real-time data, and simulated data to assess the extraction quality/accuracy and the execution speed.

3.2 Case Study: Understanding Lifestyles in Distinct Cities with Social Media Data Analysis

This study investigates the lifestyle behaviors of people living in different cities, different in size but nearby in location. Instead of traditional survey methods such as questionnaire or interviews, data from publicly available social media website is mined to analyze the desired behavior. New York City is used as a representative of large cities whereas the data from Rochester area is used as the representative of the smaller cities in the United States. To mine the prominent mobility and work-rest patterns from both the cities matrix factor analysis are used as an unsupervised method. The identified patterns are then used to quantitatively compare lifestyles of both cities.

In contrast to traditional research investigating lifestyle patterns, where data collection methods include questionnaires and telephone interviewing, we leverage data from social media to make inferences about people's lifestyles.

Among all available Location-Based Social Networks (LBSNs), Foursquare is most popular. It holds around 5 billion check-in records of about 55 million users across the world. This offers us a rich data source for conducting mobility, behavior and lifestyle studies. The study works on time and space aspects of humans and lifestyles are not assigned to any person rather non-negative matrix factorization (NMF) technique is used to discover time-based activity patterns. Time-based activities from any person's lifestyle are assumed to be correlated with his daily work-rest ratio.

To describe lifestyles according to locational behavior a spatial dimension is used. Just for an example, a primitive lifestyle pattern is defined by frequent visits to Point of Interests (POIs) such as Music venues, theaters and may be bars, while another is defined by visits to museums, art galleries, and historical locations. The clustering method is then applied to identify the behaviors of a group of individuals (e.g., Travelers or students). Additionally, third-order tensor decomposition is used to find composite patterns across both spatial and temporal dimensions. This method is an efficient way of extracting complicated patterns in multiple-dimensional spaces.

The stacked plot in Fig. 4a, b shows the check-in percentages in most popular categories (locations) during weekdays and weekends respectively in Rochester Area; whereas Fig. 5a, b represents the same for New York City. The main contribution of this study is:

- Social media data is mined as opposed to previous survey-based researches.
- Matrix Factor analysis is used to identify work-rest patterns of humans.
- Application of CP tensor decomposition to identify composite time and space lifestyle patterns.

There are many studies conducted to confirm the common perceptions of life in small cites and as well as in big cities. Just for example, life in small cities is peaceful and home-oriented and in contract, big cities' life is fast and work-oriented.

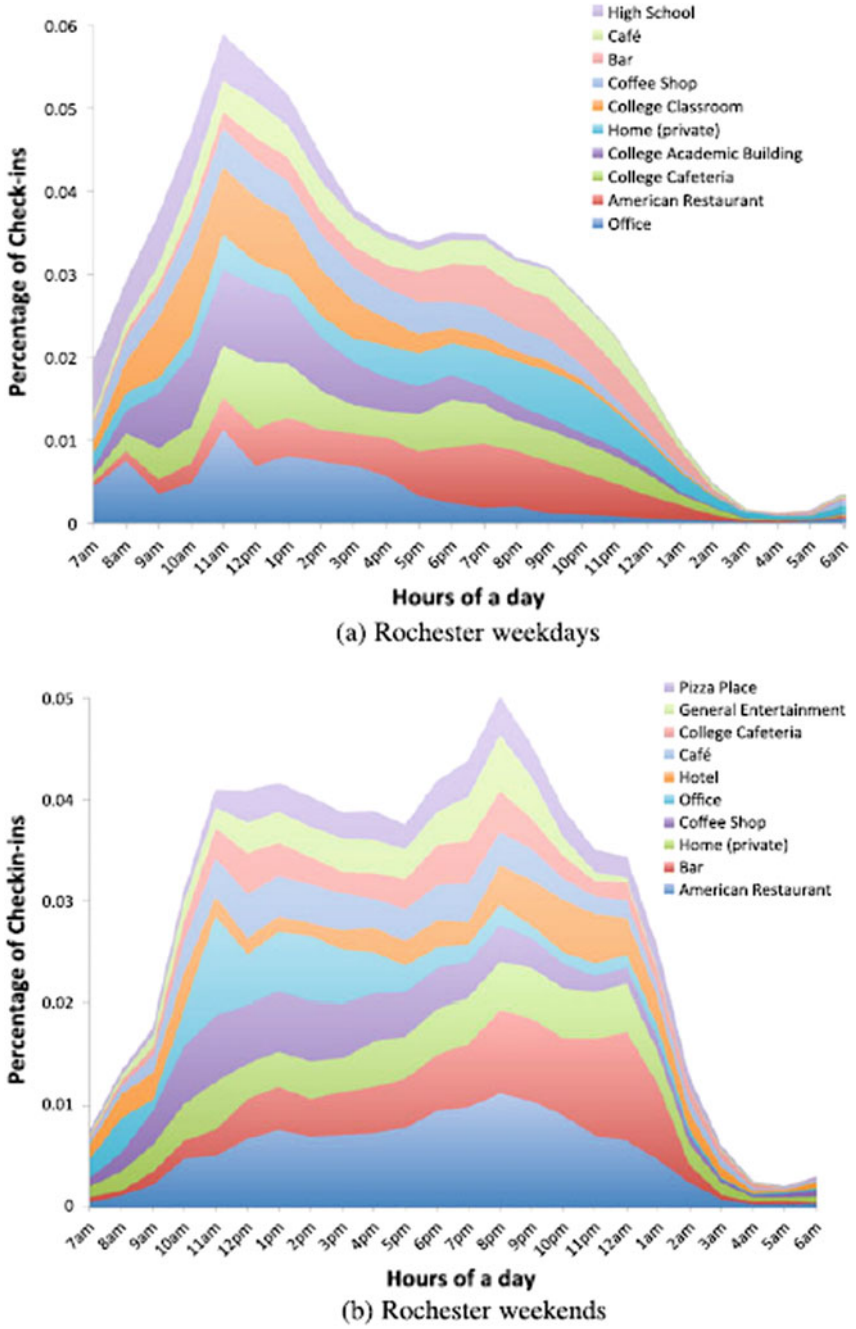


Fig. 4 10 most popular categories in Rochester, weekdays and weekends [15]

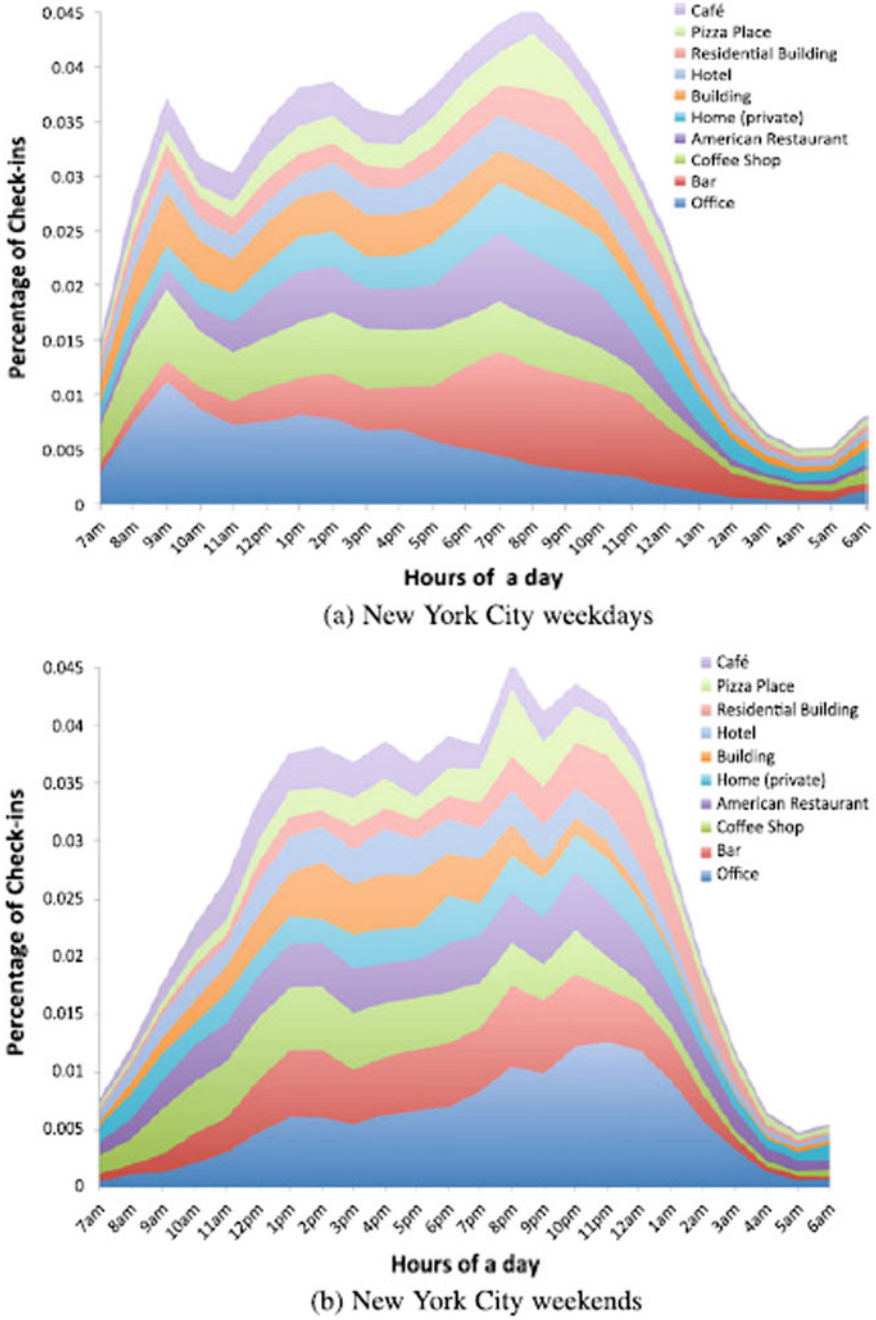


Fig. 5 10 most popular categories in New York City, weekdays and weekends [15]

4 Applications of Big Data Analytics

Analytic is said to be a set of emerged and advanced techniques and available technologies that require evaluation in integration process to disclose any type of large hidden values within any big datasets which might be specific from the standard ones, more complicated, and of a big huge scale [16].

The essential purpose of the massive records analytic is to help the company to make a higher business choice, destiny prediction; analysis massive numbers of transactions that carried out in enterprise and update the form of statistics that organization is used. Every arena of enterprise, fitness or fashionable dwelling requirements now can put in force massive information analytics.

4.1 Big Data in Healthcare

Healthcare, it is a critical area in which massive statistics analysis imposes the maximum social effect. From the analysis of capacity fitness risks in each gender to the most complex scientific analysis and calculation, massive records are present in all components of it [17]. Devices together with the Fitbit [17], Jawbone [18] and the Samsung tools match [19] permit the user to tune and upload statistics. Quickly enough such statistics will be compiled and made to be had to medical doctors if you want to aid them in the analysis.

4.2 Big Data and the World of Finance

Analyzing big data can be a very beneficial for fairly complicated inventory marketplace actions and useful resource in making worldwide economic decisions [19].

In general, as it is understood that Finance and financial system has a very large scope so it needs to be revolutionized. Numerous monetary establishments are adopting large information guidelines in order to benefit a competitive area. Complicated algorithms are being evolved to execute trades via all of the dependent and unstructured statistics won from the assets.

4.3 Big Data in Social Media Analytics

The terminology of Social Media analytics is explicitly said for the gathering data and other records form the blogs or social media websites. In this era Social Media is considered as the golden platform to apprehend the online purchaser desire or intentions and sentiments, the usage of commercial enterprise advertising over social

sites, product advertising and marketing without difficulty. EBay.com makes use of facts warehouses at 7.5 PB and as well as a 40 PB. They use a Hadoop cluster for seeking, client recommendations, and vending. Amazon.com handles thousands and thousands of back-stop operations on daily basis, as well as different queries from extra than half of one million 0.33-birthday celebration sellers [19, 20].

4.4 Big Data for the Telecom Industry

Big statistics and device mastering ideas and techniques are implemented to achieve customers' satisfaction. Element records are called, internet and client carrier logs, and emails to social media as nicely as geospatial and weather facts, there are the only few examples of information being handy to telecom service providers [21].

5 Data Visualization Using Analytical Tools

There are various tools available those can be used in many forms of evaluation techniques to maintain, control and discover meaningful inference from provided facts sets. A few of them additionally provide higher visualization by generating evaluation for summarization [22]. Records evaluation tools facilitates in deriving accurate effects with minimal efforts. This section is going to present a number of the top software suits used for facts assessment in extraordinary business organization domain names.

5.1 Data Wrangler

Wrangler [22] is designed to speed up the facts manipulation process. It enables users to spend much less time reworking on information, hence providing more time learning from it. Wrangler provides interactive transformation of cluttered, real-world data into the well-organized information tables which in turns permits spending a lot much less time in formatting and more time for analysis. Stanford University's Visualization group introduced this as a web based service. The basic idea behind its design is for cleaning and rearranging data to make it useful for other tools like spreadsheet application.

5.2 The R Project

For statistical computing and photos R programming environment [23] is available which is pretty loose environment. It is a new statistical platform that runs and accepts inputs using commands over command line. This platform has ability to provide various ways to obtain sizeable deviate, medians, correlate add-accessories and many more which allows linear models, generalized linear models time collection evaluation, nonlinear regression models, nonparametric check and classical parametric check as well as soothing and clustering. R platform can also plot charts and graphs of results. R encompasses some excellent facilities of visualization accessories along with spatial and numerical evaluation.

5.3 Tableau Plateau

Data visualization gear [23] enables absolutely everyone to put together and gift statistics intuitively. It's distinctly effective in the commercial enterprise due to it communicates insights thru data visualization. This device will flip statistics into any type of visualization accessories, from smooth to enhance.

The facility of calling bundle to counsel any visualization, drag and drop, customization of every aspect from energy recommendations and labels to size, legend display and interactive filters. It provides quite a number the way to expose interactive statistics [23]. It can be used to blend multiple linked visualization add-ons onto one dashboard, anyplace one seeks filter will act on various graphs, maps and charts; underlying tables of will also be joined.

5.4 MINITAB Software

Majorly, Minitab is a tool for statistical use and is highly used in six sigma and quality improvement of the rugged dataset. The dimensionality reduction technique used in Minitab software was Principal Component Analysis (PCA) and a Multiple Linear Regression model was fitted to see the dependency of the dependent variables on the independent variables.

5.5 SPSS

The last software which was taken into consideration for the current research was SPSS, statistical software developed by IBM. However, SPSS was not able to handle the data used for this analysis as it can accept only a limited number of rows unless

external memory is inserted for further analysis [24]. One of the major reasons behind IBM introducing new software such as IBM Watson analytics (discussed above) is because its previous software such as SPSS is not able to handle big data with given computer specification (IBM). To analyze huge amount of data the extra memory needs to be installed in the computer.

6 Conclusion

Data technological know-how alludes to a developing region of work worried about the accumulation, association, exam, perception, administration, and safeguarding of huge accumulations of statistics. Google traits and other IT fever graphs charge statistics technological know-how a number of the maximum fast growing and promising fields that grow around software program engineering. notwithstanding the fact that data technology attracts on content from set up fields like automatic reasoning, measurements, databases, belief and some greater, industry is soliciting for prepared facts researchers that no person appears to be equipped to convey.

On this chapter, we mentioned thoughts like massive statistics, information analytics, and a few fluctuated contraptions that perform statistics examination, cleansing and introduction. Big records deliver limitlessly powerful supporting strategies to the accumulation of informational indexes that's excessively thoughts boggling and considerable. The pondered contraptions spare the time spent on coding and testing with the aid of giving changed and specific results. These contraptions can be utilized as part of distinctive fields wherein facts research is needed. Statistics exam contraptions anticipate a vital component in all enterprise areas. The phase might make a contribution within the space of statistics Sciences, big statistics Analytics, and smart packages in the city, Academia, and IoT.

References

1. Donoho, D. (2015, Sept 18). 50 years of Data Science.
2. Berthold, M. R., Borgelt, C., Höppner, F., & Klawonn, F. Data Understanding.
3. https://tdwi.org/~/media/64511A895D86457E964174EDC5C4C7B1.PDF.
4. http://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science.
5. Saltz, J. S., Shamshurin, I., & Crowston, K. (2015). Comparing data science project management methodologies via a controlled experiment. https://www.nap.edu/read/23670/chapter/6.
6. Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493–1507. http://dx.doi.org/10.5465/amj.2016.4005.
7. http://www.saedsayad.com/model_deployment.htm.
8. http://analyticscanvas.com/knowledge-base/what-is-data-preparation/.
9. Godbole, N. S., & Lamb, J. (2015). Using data science big data analytics to make healthcare green. In *2015 12th International Conference Expo on Emerging Technologies for a Smarter World (CEWIT)*.

10. Ahmad, M. O., Markkula, J., & Oivo, M. (2013). Kanban in software development: A systematic literature review. In *2013 39th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA)* (pp. 9–16). IEEE.
11. Hayes, M. A., & Capretz, M. A. M. (2015, Feb). Contextual anomaly detection framework for big sensor data. *Journal of Big Data, 2*, 2.
12. Kenyon, D., & Eloff, J. H. P. (2017). Big data science for predicting insurance claims fraud. In *2017 Information Security for South Africa (ISSA)*.
13. Hu, T., Bigelow, E., Luo, J., & Kautz, H. (2017, March). Tales of two cities: using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. *IEEE Transactions on Big Data, 3*, 55–66.
14. Dobrkovic, M.-E. I., & van Hillegersberg, J. (2018, January). Maritime pattern extraction and route reconstruction from incomplete AIS data. *International Journal of Data Science and Analytics*.
15. Barrachina, D., & OḎriscoll, A. (2014, June). A big data methodology for categorizing technical support requests using Hadoop and Mahout. *Journal of Big Data, 1*, 1.
16. Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques, and technologies: A survey on big datal, contents lists available at *Science Direct Information Sciences, 275*, 314–347.
17. Kaisler, S., Frank Armour, J., Espinosa, A., & Money, W. (2013). *Big data: issues and challenges moving forwardl, 2013 46th Hawaii International Conference on System Science 1530-1605/12*, *2012 IEEE*.
18. http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/.
19. Web content available on the link: Retrieved August 16, 2015, from http://searchbusinessanalytics.techtarget.com/definition/social-media-analytics.
20. Web content available on the link: Retrieved August 16, 2015, from http://en.wikipedia.org/wiki/Ecommerce_in_India#cite_noteOnline_shopping_touched_new_heights_in_India_in_2012-1.
21. Yan, J. (2013, April 9). Big Data, Bigger Opportunities Bowman, M., Debray, S. K., & Peterson, L. L. (1993). Reasoning about naming systems.
22. Research in big data and analytics: An overview. *International Journal of Computer Applications (097 –8887), 108*(14) (2014, December).
23. Chen, P. C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques, and technologies: A survey on big data. *Information Sciences, 275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015.
24. George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal. Academy of Management, 57*(2), 321–326. https://doi.org/10.5465/amj.2014.4002.

Data Science and Digital Business

Lei Bu and Feng Wang

List of Abbreviations

GIS	Geographic Information System
NAVD88	North American Vertical Datum
SLOSH	Sea, Lake and Overland Surges from Hurricanes
NOAA	National Oceanic and Atmospheric Administration
FEMA	Federal Emergency Management Agency
NHC	National Hurricane Center
HPS	Hurricane Protection System
ADT	Average Daily Traffic
DEM	Digital Elevation Model
MDOT	Mississippi Department of Transportation
CI	Condition Index
VIF	Variance Inflator Factor
UTC	University Transportation Center

L. Bu
Department of Civil and Environmental Engineering, Institute for Multimodal Transportation, Jackson State University, 1400 Lynch Street, PO Box 17068, Jackson, MS 39217-0168, USA
e-mail: leibu04168@gmail.com

F. Wang (✉)
Ingram School of Engineering, Texas State University, 601 University Dr., San Marcos, TX 78666, USA
e-mail: f_w34@txstate.edu

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_2

1 Introduction

Hurricanes are among the most catastrophic events resulting in severe consequences including loss of lives and property damage. A hurricane produces storm surges, tornadoes, and inland flooding, while storm surge is always a potential threat to the coast area. The Mississippi Gulf Coast area traditionally refers to the three Mississippi coastal counties namely Hancock County, Harrison County, and Jackson County that lie along the Gulf of Mexico. The residential properties and road networks in this area are constantly vulnerable to hurricanes as well as floods and severe thunderstorms. The magnitude of devastation was quite evident in hurricanes Katrina and Rita in the Mississippi Gulf Coast [1]. The destructive power of the 2005 Hurricane Katrina overwhelmed the Mississippi coastline and the Gulf Coast of Mississippi which suffered nearly total devastation with the strong hurricane winds, extreme storm surge, and waves pushing casino barges, boats and debris into towns, and leaving 236 people dead, 67 missing, and an estimated $125 billion dollars in damages [2]. Numerous streets and bridges were washed away, including the bridge sections of highway US 90. In particular, the roadway/bridge portion of US 90 between Bay St. Louis and Pass Christian was completely destroyed by the storm surge leaving only the supporting structure [2].

In addition, the population is firmly increasing in coastal areas as over half of the Nation's economic productivity is located within a narrow coastal belt. From 1990 to 2008, the population density increased by 32% in the Gulf coastal counties [3, 4] making them the most densely populated counties in the United States. It is also a fact that the Gulf Coast coastlines lie less than 10 feet above NAVD83 datum (North American Vertical Datum). Literally 72% of ports, 27% of major roads, and 9% of rail lines within the Gulf Coast region are at or below 4 feet elevation. Consequently, a storm surge of 23 feet has the ability to inundate 67% of interstates, 57% of arterials, almost half of rail roads, 29 airports, and virtually all ports in the Gulf Coast area [4]. Hence, it is necessary to analyze the risk of hurricane induced flooding in this area.

This study aims to visualize and analyze the flooding risk of the infrastructure and properties of the Mississippi Gulf Coast area due to storm surge of a potential hurricane and to possibly provide more resiliency to the transportation system under extreme weather conditions. Clearly it is not only necessary to develop a collection of proprietary datasets for the flooding vulnerability characteristics of the area, but it is also imperative to identify the critical risk related factors and quantify the vulnerability risk for the emergency management officials and planners to be proactive about the risk and be prepared for deploying effective emergency management strategies for the communities, residential properties, and public infrastructure including the transportation systems of the area. The objectives of this study are: (1) to retrieve and visualize the data items relevant to the storm surge induced flooding risk in the Mississippi Gulf Coast area; and (2) to analyze the variables significantly related to the storm surge induced flooding risk in the area.

The book chapter is organized as follows: after the Introduction section, the Background section reviews various studies on flooding risks and related issues using analysis and visualization technologies, and storm surge models. The methods used for this study are summarized in the Study Methodology section. In the Data Collection and Retrieving section, the surge flooding related data are described and presented for the study area. In the Risk Analysis and Prediction section the statistical analysis results and regression models are presented. Finally, findings and observations of the study are summarized in Conclusions.

2 Background

The characterization of the impacts of storm surge and inundation due to a hurricane on the residential properties, traffic networks, and critical infrastructure, and predictive risk analysis of the impacts of the storm surge induced flooding have been a focus of the recent hurricane studies.

2.1 Modeling of Storm Surge and Inundation

Prior research has been conducted to describe how an extreme climate affects the critical transportation infrastructure in storm surge and inundation using general circulation models of the atmosphere and the Geographic Information System (GIS) presentation tools. Bates et al. [5] in UK coded on the LISFLOOD-FP and GIS tools to show a simplified two dimensional model for coastal flooding. In the study, Bates et al. observed the maximum inundation extent for particular events and predicted the coastal inundation using GIS data. In 2013 Bates released the 5.9.6 version of the LISFLOOD-FP program which is a two-dimensional hydrodynamic model specifically designed to simulate floodplain inundation in a computationally efficient manner over complex topography [6]. In the U.S., the Sea, Lake and Overland Surges from Hurricanes (SLOSH) of the National Oceanic and Atmospheric Administration (NOAA) is capable of predicting future sea level rise and storm surge height for a coastal area in the U.S. In developing the SLOSH, scatter plots and probability density functions of average temperature and precipitation change were derived from the general circulation model and the models used the historical tracking and meteorological data of dated North Atlantic tropical storms from 1851 to present [7]. This study will collect storm surge data from NOAA SLOSH SDP based on NOAA SLOSH model to determine the inundation height.

Hurricane surge threat under climate change was examined by Lin et al. [8] while simulating cyclone surges for different climates. A practical approach was taken to couple a simpler general circulation model such as a GCM-driven (Global Climate model) statistical/deterministic hurricane model with hydrodynamic surge models. The authors were able to identify that certain climate models increase storm surge

flooding due to a change in storm sizes. Recently, GIS was applied by Dr. Kent [9] to estimate the storm surge inundation affecting Louisiana. In the study, the author synthesized the hazard data associated with flooding and found that the flood risk of a roadway depends on the road class and the geographical and environment conditions. GIS spatial statistics analysis methods will be used in this study to analyze traffic, population, storm surge, inundation, and loss coverage.

Savonis et al. [10] studied the central Gulf Coast transportation network in relation to storm activity. Most of the spatial data was organized in GIS formats that can be integrated to assess the vulnerability and risks of the transportation infrastructure in the study area and inform the development of adaptation strategies. In the study, it indicated that Category 3 and higher intensity level storms may return more frequently to the central Gulf Coast and thus cause more disruptions to the transportation networks. Rising relative sea levels would also exacerbated exposure of the area to storm surge and flooding. They also found that depending on the trajectory and scale of individual storms, facilities at or below 9 m (30 feet) could be subject to direct storm surge impacts. Hyman et al. [11] addressed the risk that a climate change poses for the state, region, and the local planning of transportation organizations. To summarize these researches, the Gulf Coast region had austere and realistic needs for dealing with vulnerability risks of surge flooding due to climate change and the invasion of a potential hurricane to better protect the very dense populations and the social-economically important multi-modal transportation systems in the area.

2.2 Risk Analyses in Hurricane Studies

Since Baker [12] early studied the hurricane evacuation behavior in risk prone areas along the gulf coast region. Five variables including action by public authorities, housing, prior perception of personal risk, and storm-specific threat factors to identify evacuation rates risk level were used. Humphrey [13] identified climate change factors that are particularly important to the transportation system in the United States including rising sea levels, increases in intense precipitation, and increases in hurricane intensity. Ayyub et al. [14] analyzed social vulnerability examined the resulting surge, waves, and precipitation to calculate the performance of a hurricane protection system. In that study the Hurricane Protection System was (HPS) such as levees, floodwalls, pumping station, were considered. They later concluded their findings by evaluating the usage of hurricane protection system and its contribution towards the population and properties. Lu et al. [15] explored an accessibility-based criticality prioritization methodology to identify and prioritize critical transportation infrastructure. The methodology was applied to the road network of Hillsborough County Florida, which was threatened by flood risk from storm surge, sea-level rise, and intense precipitation. The vulnerability of population during hurricanes was studied by Bian and Wilmot [16] by using the Geographic Information System. Six type of population group were considered in order to capture the different effects to the population during evacuation. Study found that the population, land elevation, and traffic are key components when dealing with risk due to flooding.

Myers et al. [17] investigated place-based social vulnerability and post-disaster migration in the U.S. Gulf coast region. They analyzed the demographic, social, and the economic data of places that were affected by hurricanes Katrina and Rita. A regression analysis was made between migration and social vulnerability. Choate et al. [18] discussed the methodology and results of a broad vulnerability assessment of the highly critical transportation assets in Mobile in Gulf Coast. Key components of vulnerability was analyzed and indicator-based vulnerability assessment was conducted for representative highway segments, rail segments, and pipeline segments.

The methods presented in past studies such as GIS analysis, general circulation method, factor analysis, and regression analysis especially in Gulf Coast are relevant and helpful to the study. This study will address the analyses and prediction of storm surge induced flood risk in Gulf Coast area of Mississippi.

3 Methodology

3.1 Study Area Description

Gulfport is located at the center of the Mississippi coast and it is the second largest city in Mississippi after the state capital Jackson.

It is about 70 miles of highway distance from New Orleans, LA. High freight traffic is generated due to the transshipping of freight from cargo vessels to trucks and intermodal operations at Port of Gulfport, which is No. 19 in the U.S. in terms of containership and is among the top 50 U.S. ports by port calls and vessel type [3, 19]. The majority (88%) of the populations of Mississippi Gulf Coast are located within the three coastal counties which are: Hancock County, Harrison County, and Jackson County. These counties are included in the study.

The study area is shown in Fig. 1a [1]. According to U.S. census data for July 1, 2015 [3], the population in Hancock County, Harrison County and Jackson County is 46,420, 201,410 and 141,425 respectively. Global peak surge map from SURGEDAT [20] in Fig. 1b shows that the peak surge along the Gulf Coast is equal to or larger than 4.5 m (15ft) which represents severe flooding.

3.2 Datasets

In this study data preparation, spatial analysis method using GIS, and statistical analysis method using SAS were used to show the flooding risk along the Gulf Coast area. Data of population census, average daily traffic, elevation of the land surface, and direct loss coverage were collected from U.S. Census Bureau, Mississippi Department of Transportation, USGS, and Mississippi Insurance Department, respectively. Storm surge height above the NAVD88 datum was collected from NOAA the Sea, Lake and Overland Surges from Hurricanes (SLOSH) Display Program (SDP) and historical hurricane storm surge data collected by FEMA [21] and National Hur-

(a) Map of Gulf Coast area

(b) Global peak surge map

Fig. 1 Maps of the study area

ricane Center (NHC) [22]. Hurricane inundation level above the land surface was calculated by the formula proposed considering the method form NOAA. GIS technology was used to conduct data management. Maximum flood inundation heights above the land surface are calculated based on the elevation of the land surface subtracting from the surge elevation. GIS Spatial analyst tool was used to obtain the data distribution results in census blocks including direct loss coverage. Later statistical method was used for risk factor analysis and prediction for direct loss coverage.

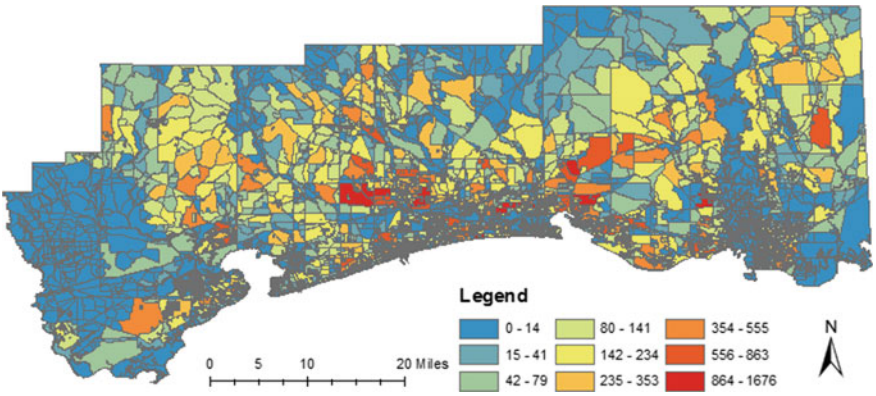
4 Risk Related Data and Visualization

4.1 Population and Traffic Data

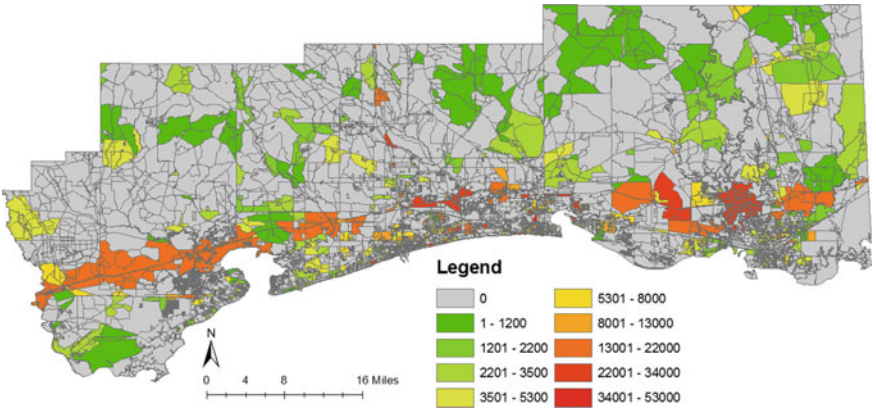
Population and traffic data collected are shown in Fig. 2. Census block data in Gulf Coast area was collected from U.S. Census Bureau [3].

Figure 2a shows the population census data for 2015. There are 14,318 census blocks in the area. The deep red color represents the highest density of population and the deep blue color represents the lowest density of population. The higher density of population in Harrison County and Jackson County is noticeable.

Average Daily Traffic (ADT) data was collected from Mississippi Department of Transportation (MDOT). Each block is given a summation of traffic values for all the road segments which intersect the block. Figure 2b shows the Average Daily Traffic (ADT) blocks data in 2015. The orange color and red color on east-west (e.g. Interstate 10) and south-north (e.g. highway US 49) represents the higher traffic volume.



(a) Population census data in 2015



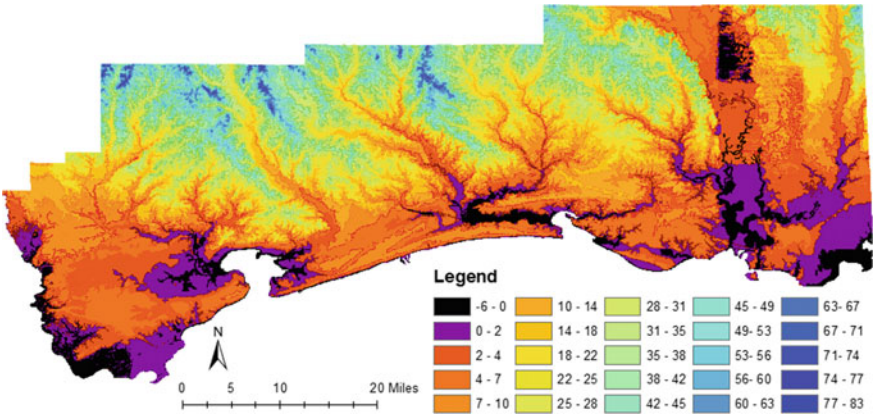
(b) ADT traffic data in census blocks in 2015

Fig. 2 Population and traffic data in the Gulf Coast (Color figure online)

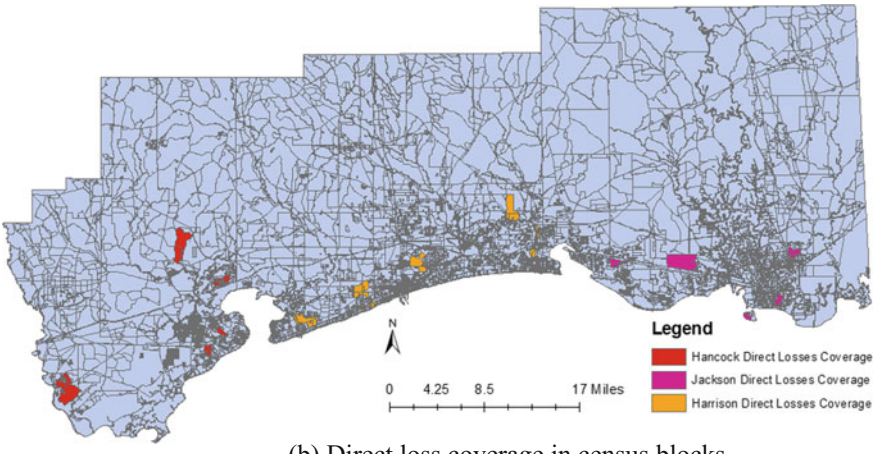
4.2 *Elevation and Direct Loss Coverage Data*

Elevation and direct loss coverage data collected are shown in Fig. 3.

Digital Elevation Model (DEM) at 5 ft horizontal resolutions was collected from USGS [23] based on horizontal datum NAD83 and vertical datum NAVD88. The data was imported into grid format by ArcMap. Figure 3a shows the elevation data in the area. Different color represents different elevation values. The dark color in the figure represents the lower elevation below the datum. The deep blue color represents the higher elevation above the datum. Direct loss refers to the loss incurred due to direct damage to property. According to the data [24], Hurricane Katrina in 2005 thwarted all the other loss events over the last 11 years. Direct losses in dollar incurred due



(a) Elevation data in the area



(b) Direct loss coverage in census blocks

Fig. 3 Elevation and direct loss coverage in the Gulf Coast (Color figure online)

to windstorm covered by zip codes and cities in each county from year 2005 to year 2007 were collected from Mississippi Insurance Department. Average direct loss coverage for blocks in each county are shown in Fig. 3b.

4.3 Hurricane Storm Surge Data

Storm surge, tides, and waves are key factors contributing to coastal hurricane floods and severe damage to the resident and facilities in the coastal counties [9, 22].

First, strong winds and low atmospheric pressure caused by tropical cyclones such as hurricanes could drive up the water level to create a storm surge. Second, high tide levels are caused by normal variations in the astronomical tide cycle. When a severe storm hits during high tide, the risk of flooding increases significantly. Last, large waves driven by local winds or swelled from distant storms could raise average coastal water levels and cause large and damaging waves to reach land [22].

Storm surge height data of 2005 Katrina above NAVD88 in MS Gulf coast basin was collected from NOAA SLOSH SDP. Fig. 4 shows the SLOSH storm surge model for the MS Gulf Coast basin, and storm surge height in feet at a sample location is presented in Fig. 5a.

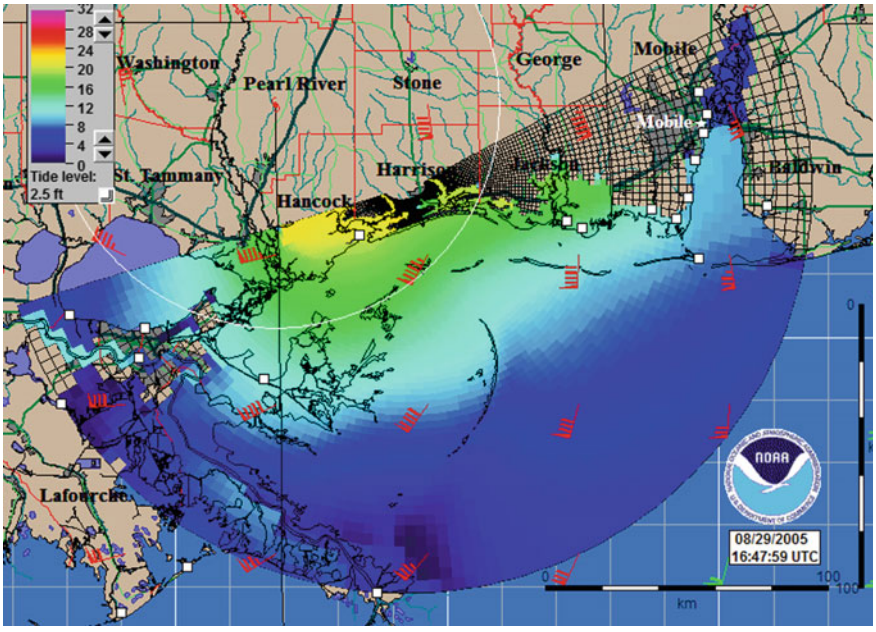
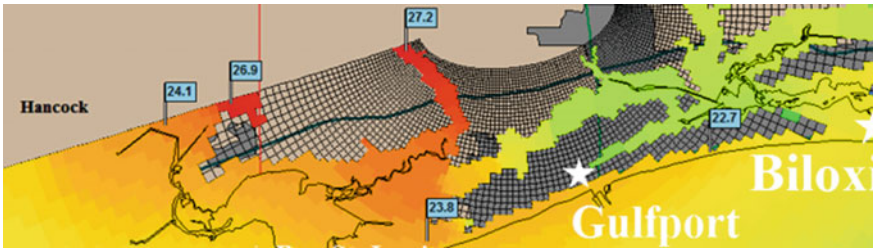
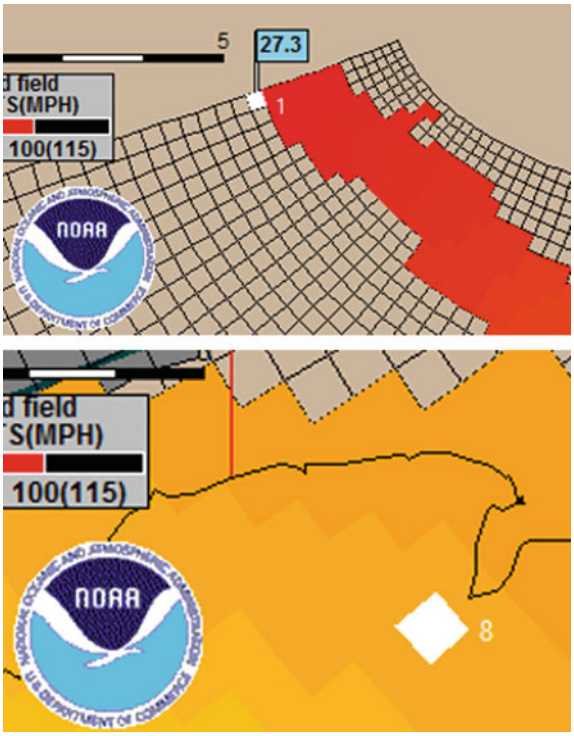


Fig. 4 SLOSH storm surge model for the Gulf Coast, Mississippi basin



(a) Storm surge at different locations



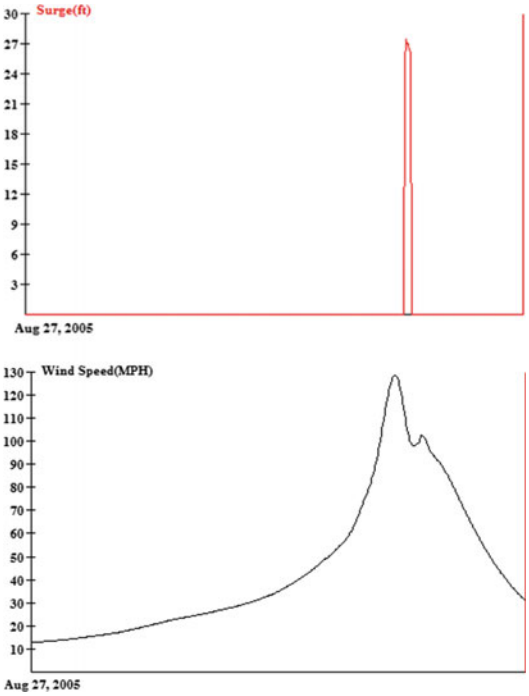
(b) Locations selected

Fig. 5 Storm surge results at different locations

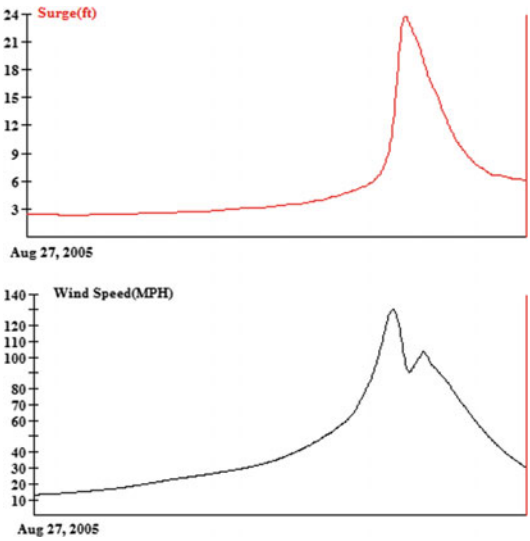
As sample results, location 1 on the basin boundary and location 2 in the basin are selected respectively, as shown in Fig. 5b to show surge height and wind speed in the time period studied which is shown in Fig. 6.

The maximum surge height in the time period is collected for all the locations to be the storm surge data in this study which is partly shown in Fig. 7.

Figure 7a, b and c show the storm surge height over 3.5 m (11.5 feet) above NAVD 88 datum in Harrison County, Hancock County, and Jackson County respectively.

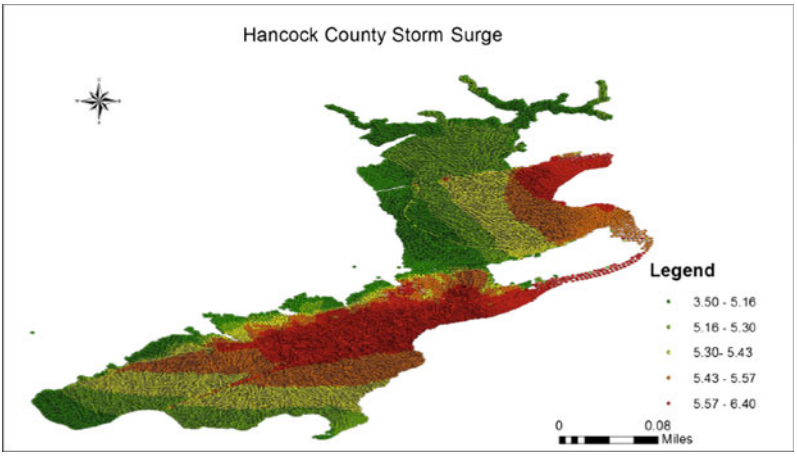


(a) Surge height and wind speed data at location 1

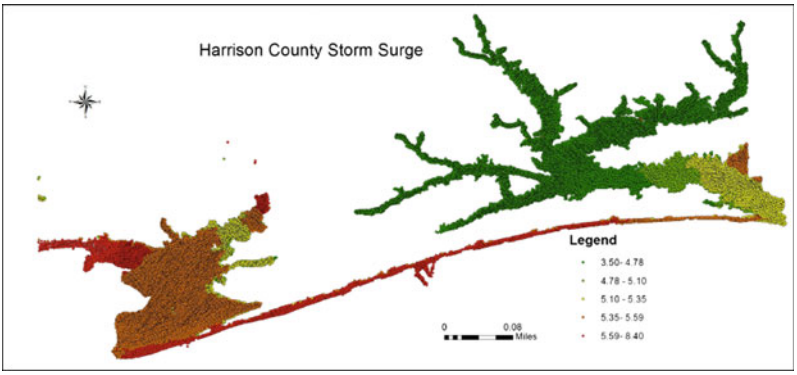


(b) Surge height and wind speed data at location 8

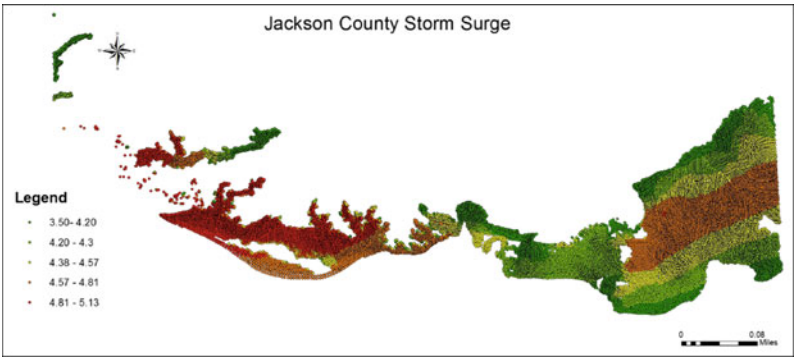
Fig. 6 Surge height and wind speed in the time period studied



(a) Hancock County



(b) Harrison County



(c) Jackson County

Fig. 7 Storm surge data collected in Gulf Coast area

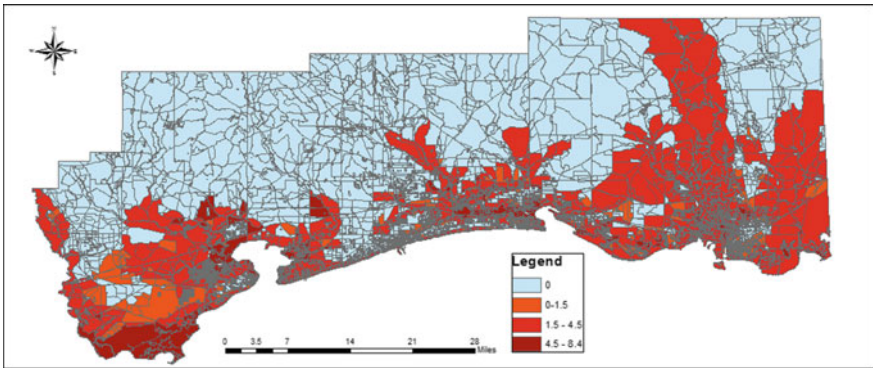


Fig. 8 Flood inundation height over the land surface

There are much more risk in Harrison County and Hancock County by showing the more higher storm surge from 5.5 m (18 feet) to 8.4 m (27.5 feet) than that in Jackson County with lower storm surge less than 5.5 m (18 feet).

4.4 Flood Inundation

Based on the computerized model developed by the National Weather Service [25], Flood inundation height over the land surface is calculated by formulation (1).

$$H_l^i = H_w^i - H_t - E_l^i, i = 1, 2, \ldots n \tag{1}$$

where, H_l^i is the flood inundation height over land surface of location i, H_w^i is the maximum estimated storm surge height of location i, E_l^i is the land surface elevation of location i, H_t is tide level of MS Gulf Coast basin, and n is the total number of locations. The average flood inundation height in blocks is presented in Fig. 8. The red and pink color show that the inundation height is larger than 4.5 m. There are higher inundation both in Hancock County and in Harrison County where elevation is relatively low.

5 Risk Analyses and Prediction

There are four variables: average direct loss (Avg_L), population (POP), average daily traffic (ADT), and average flood inundation height (Avg_H). Dependent variable is average direct loss. Independent variables are population, average daily traffic, and average flooding inundation height. The regression model is described as follows:

$$Avg\_L = \beta_1 \cdot POP + \beta_2 \cdot ADT + \beta_3 \cdot Avg\_H + c \quad (2)$$

where parameters β_1, β_2, and β_3 are coefficients of dependent variable population, average daily traffic, and flood inundation height, respectively. The collinearity analysis, and regression analysis are tested using SAS.

5.1 Collinearity Analysis and Multiple Regression Analysis

Collinearity among the independent variables in the regression model was tested and the results of eigenvalue, Condition Index (CI), tolerance, and Variance Inflator Factor (VIF) are shown in Table 1.

The outputs of Condition Index (CI) was maximum 3.3466 and the Variance Inflator Factor (VIF) was $1 < VIF < 5$. According to statistical studies, if the Condition Index is less than 30, there is no significant collinearity problem [26]. In addition, there is a significant collinearity problem when Variance Inflator Factor is 10 or above [27]. It was concluded that there was no collinearity problem with this regression model.

Multiple Regression analysis was tested in SAS to show the parameter estimate, t-value, and P-value in Table 2.

The significant low P-value (P-value < 0.01) indicates that the null hypothesis is rejected. If the population has no effect on the direct loss coverage, we would obtain the observed difference or more in 0.93% of analysis due to random sampling error. If average daily traffic has no effect on the direct loss coverage, we would obtain the observed difference or more in 0.08% of analysis due to random sampling error. And if average inundation height has no effect on the direct loss coverage, we would obtain the observed difference or more in 0.01% of analysis due to random sampling

Table 1 Collinearity analysis results

Model	Eigenvalue	CI	Tolerance	VIF
POP	0.4786	2.4525	0.9391	1.0649
ADT	0.3859	2.7311	0.8322	1.2017
Avg_H	0.2570	3.3466	0.8596	1.1633

Table 2 Multiple regression analysis results

Model	Parameter estimate	t-value	Pr>|t|
C	7,274,675	–	–
POP	81,182	2.68	0.0093
ADT	−943.0887	−3.51	0.0008
Avg_H	14,097,139	5.36	<0.0001
R	0.877		
R square	0.769		

error. R square is 0.769 showing that the model explains 76.9% variability of the average direct loss coverage around its mean.

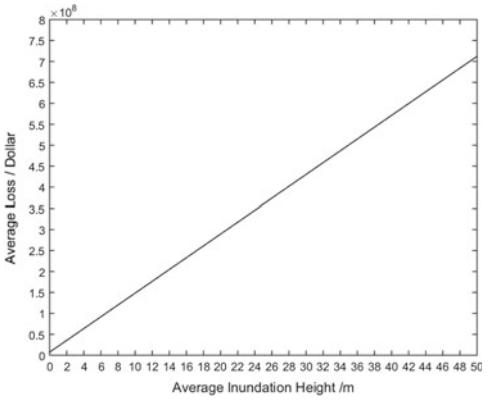
We notice that the direct loss will increase as population or average inundation height increases because there is residential population in the storm surge induced flood risk area and flood inundation causes the direct loss for both of the population and facility in the area. But the direct loss will decrease as average daily traffic increases mainly because the heavy traffic in the area is on Interstate 10 and highway US 49 respectively, which are not in high flooding risk with storm surge. The average daily traffic increases implies that more proportion of population are transferred to these two highways and decrease the risk affected.

5.2 Risk Prediction

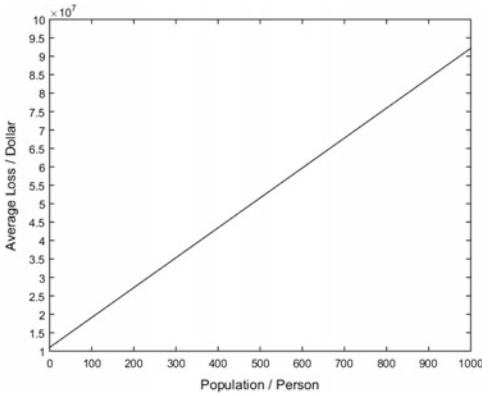
Based on the regression model, risk prediction was conducted to show how this study will contribute to the risk prediction, planning and management. A sample census block in the study area with all the features needed was selected and the risk prediction results are shown in Fig. 9.

There are three predictions including: (1) Fixing population and annual daily traffic in the block to find the average direct loss values as the flood inundation height continuously increases from 0 to 50 m as shown in Fig. 9a; (2) Fixing average inundation height and average daily traffic in the block to find the average direct loss values as the population continuously increases from 0 person to 1000 persons as shown in Fig. 9b; and (3) fixing population and average inundation height in the block to find the average direct loss values as the average daily traffic continuously increases from 0 vehicle to 50,000 vehicles as shown in Fig. 9c.

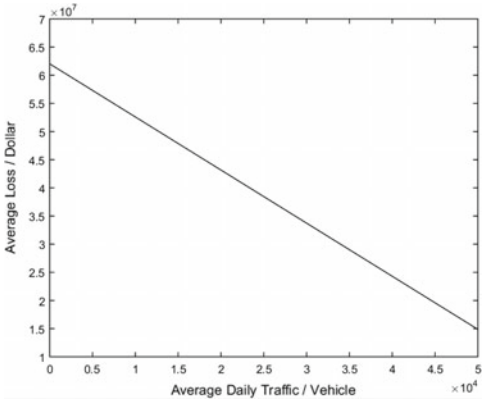
Predict results show that the average direct loss increases as flood inundation height or population continuously increases, while the average direct loss decreases as average daily traffic continuously increases. These will provide emergency management strategies for emergency responses in hurricane to decrease the direct loss in a hurricane. For example, constructing or improving the coastal structure to decrease flood inundation height, evacuating residents ahead of the hurricane storm sure induced flood risk to decrease the population exposed in the risk area, and facilitating evacuation by highway routes with increased road capacity to promote the resident and vehicle evacuation, etc. This prediction method could be used for other sample blocks in the study area to provide necessary emergency information for the emergency responses.



(a) Prediction with fixed POP and ADT



(b) Prediction with fixed Avg_H and ADT



(c) Prediction with fixed POP and Avg_H

Fig. 9 Risk prediction results for a block case

6 Conclusions

This book chapter analyzes and predicts flooding risk from hurricanes along the Gulf coast area using Geographic Information System and multiple regression analysis by SAS. Data of population census, average daily traffic, storm surge, and surface elevation, and direct loss coverage was collected and managed in census block. Maximum flood inundation height above the land surface was calculated based on storm surge and tide level in the study area. A regression model was obtained by testing multicollinearity, and multiple regression analyses to effectively predict the direct loss under the different conditions. Based on the study results, following observations are made:

(1) There are much more risk in Harrison County and Hancock County by showing the more higher storm surge from 5.5 m (18 feet) to 8.4 m (27.5 feet);
(2) There are higher flood inundation both in Hancock County and in Harrison County where elevation is relatively low;
(3) Collinearity analysis results indicated no collinearity problem with this regression model and the model explains 76.9% variability of the average direct loss coverage around its mean; and
(4) Predict results could provide emergency management strategies for emergency responses in hurricane to decrease the direct loss in a hurricane.

This study analyzes and predicts flood risk due to hurricanes along the Gulf Coast Area based on direct loss coverage incurred which only includes windstorm coverage from year 2005 to year 2007 in hurricane Katina. How to apply this regression model or concept to a more wide hurricanes and insurance coverage will be the research work in the future.

Acknowledgements The project was partially funded by the Institute for Multimodal Transportation (IMTrans) at Jackson State University through the UTC program of the US Department of Transportation (USDOT).

References

1. Mississippi Gulf Coast. (2016). Retrieved from January, 2016, from https://en.wikipedia.org/wiki/Mississippi_Gulf_Coast.
2. Wiki. (2015). Effects of Hurricane Katrina in Mississippi. Retrieved December 1, 2015, from http://en.wikipedia.org/wiki/Effects_of_Hurricane_Katrina_in_Mississippi.
3. United States Census Bureau. (2016). Retrieved March 3, 2016, from http://www.census.gov.
4. National Hurricane Center. (2016). Retrieved February 13, 2016, from http://www.nhc.noaa.gov.
5. Bates, P. D., Dawson, R. J., Hall, J. W., et al. (2005). Simplified two-dimensional numerical modelling of coastal flooding and example applications. *Coastal Engineering, 52*(9), 793–810.
6. Bates, P., Trigg, M., Neal, J., & Dabrowa, A. (2013). *User Manual for LISFLOOD-FP Code Release 5.9.6*. University of Bristol, UK.

7. Jelesnianski, C. P., Chen, J., & Shaffer, W. A. (1992). *SLOSH: Sea, lake, and overland surges from hurricanes.* NOAA Technical Report NWS 48, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 71 pp.
8. Lin, N., Emanuel, K., Oppenheimer, M., & Vanmarcke, E. (2012). Phyically based assessment of hurricane surge threat under climate change. *Nature Climate Change, 2,* 462–467.
9. Kent, J. D. (2013, January). *Quantifying the key factors that create road flooding*. Center for GeoInformatics (C4G), Louisiana State University. LTRC Project Number: 11-6GT.
10. Savonis, M. J., Burkett, V. R., & Potter, J. R. (2008, March). Impacts of climate change and variability on transportation systems and infrastructure: Gulf Coast study, Phase I. In *U.S. Climate Change Science Program, Synthesis and Assessment Product 4.7.*
11. Hyman, R., Lupes, R., & Perlman, D. (2011). Federal highway administration activities related to the adaptation of transportation infrastructure to climate change impacts. Chapter of adapting transportation to the impacts of climate change: State of the practice 2011. *Transportation Research Board,* 12–18.
12. Baker, E. J. (1991). Hurricane evacuation behavior. *International Journal of Mass Emergencies and Disasters, 9*(2), 234–245.
13. Humphrey, N. P. (2008). TRB Special Report: Potential impacts of climate change on U.S. transportation. *Transportation Research Board, 256,* 21–24.
14. Ayyub, B. M., Foster, L., & Mcgill, W. L. (2009). Risk analysis of a protected hurricane-prone region. I: Model development. *Natural Hazards Review*, *10*(2), 38–53.
15. Lu, Q. C., Peng, Z. R., & Zhang, J. J. (2015). Identification and prioritization of critical transportation infrastructure: Case study of coastal flooding. *Journal of Transportation Engineering*, *141*(3), 04014082-1–04014082-8.
16. Bian, R. J., & Wilmot, C. G. (2016). Measuring the vulnerability of disadvantaged populations during Hurricane evacuation. In *Transportation Research Board 95th Annual Meeting,* Washington DC, 10–14 January 2016.
17. Myers, C. A., Slack, T., & Singelmann, J. (2008). Social vulnerability and migration in the wake of disaster: The case of Hurricanes Katrina and Rita. *Population and Environment, 29*(6), 271–291.
18. Choate, A., Evans, C., Rodehorst, B., Saavedra, R., Snow, C., Snyder, J., et al. (2014, June). Impacts of climate change and variability on transportation systems and infrastructure: The Gulf Coast study, phase 2. US Department of Transportation. Report No.: FHWA-HEP-14-033.
19. U.S. Department of Transportation (USDOT). (2014). Maritime Administration, U.S. Waterborne Foreign Container Trade by U.S. Custom Ports. Retrieved November 10, 2014, from https://www.marad.dot.gov/resources/data-statistics/.
20. Louisiana State University (LSU). (2016). Global Peak Surge Map. SURGEDAT. Retrieved January 15, 2016, from http://surge.srcc.lsu.edu.
21. The Federal Emergency Management Agency. (2016). Retrieved January 15, 2016, from https://www.fema.gov/.
22. National Hurricane Center. (2016). Retrieved February 4, 2016, from http://www.nhc.noaa.gov.
23. U.S. Geological Survey (USGS). (2016). Retrieved January 15, 2016, from https://www.usgs.gov/.
24. Mississippi Insurance Department. (2015, December). *Analysis and interpretation of the clarity act data call.*
25. National Weather Service. (2016). Retrieved January 17, 2016, from http://www.weather.gov/.
26. Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression* (1 st ed.). Wiley.
27. O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690.

Data Science and Conversational Interfaces: A New Revolution in Digital Business

David Griol and Zoraida Callejas

1 Introduction

Conversational interfaces have become increasingly popular during the last years supported by recent major advances in Artificial Intelligence (data science and deep learning to cite just two), language technologies (e.g., automatic speech recognition, natural language processing, and use of the semantic web), and device technologies (Internet of Things, more powerful smartphones, use of sensors and context information, and increased connectivity) [13, 14, 20]. A sign of the promise of the advances in these fields is that many major companies such as Google, Microsoft, Amazon, Facebook, Samsung or Baidu have developed and continuously improve high-profile conversational assistants such as Siri, Google Now, Alexa, Cortana and Bixby.

With the growing maturity of conversational technologies, the possibilities for integrating conversation and discourse have also been extended to a number of new human-machine interaction domains. One of the most wide-spread and initial applications of these systems is information retrieval from the web, database systems, and recommendation systems [7, 33]. More recent application domains include speech controlled telephone banking systems [15], conference help assistants [2], tutoring systems [21], question-answering applications [28], conversation practice for language learners [5], pedagogical agents and learning companions [3], interaction with robots [10], dialog applications for computer-aided speech therapy with different language pathologies [26], etc.

Natural human-computer interaction is a complex problem that requires work on multiple levels, such as speech recognition, natural language processing, dialog management, and speech synthesis. However, most of the natural language conversational interfaces are still oriented to solve restricted problems in which the knowledge

D. Griol (✉)
Computer Science Department, Universidad Carlos III de Madrid, Leganés, Avda. de la Universidad, 30, 28911 Leganés, Spain
e-mail: dgriol@inf.uc3m.es

Z. Callejas
Department of Languages and Computer Systems, University of Granada, CITIC-UGR, C/Pdta. Daniel Saucedo Aranda s/n, 18071 Granada, Spain
e-mail: zoraida@ugr.es

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_3

is stored in just one type of source (e.g., relational databases, ontologies, etc.) and hand-crafting dialog strategies tightly coupled to the application domain are still used by many commercial conversational systems to offer exactly the same system's behavior for every user.

In recent years, the computational linguistics community has focused on developing language processing algorithms that can leverage the vast amounts of data (known as Big Data [6]) that are generated every day. In this field, Data Science and Machine Learning techniques can potentially reduce human effort in the knowledge engineering process, enable these AI systems to learn and become increasingly more intelligent, adapt these interfaces considering users' specific preferences, and facilitate developing, deploying and re-deploying conversational systems.

Thus, learning statistical approaches to model the different modules that compose a conversational system has been of growing interest during the last years [8, 31]. Models of this kind have been widely used for speech recognition and also for language understanding [1, 9]. Automating dialog management is useful for reducing the time-consuming process of hand-crafted design of dialog strategies. In fact, the application of machine learning approaches to dialog management strategy design is a rapidly growing research area. Machine-learning approaches to dialog management attempt to learn optimal strategies from corpora of real human-computer dialog data using automated "trial-and-error" methods instead of relying on empirical design principles [23]. The main trend in this area is an increased use of data for automatically improving the performance of the system.

Statistical models can be trained with corpora of human-computer dialogs with the goal of explicitly modeling the variance in user behavior that can be difficult to address by means of hand-written rules [23]. Additionally, if it is necessary to satisfy certain deterministic behaviors, it is possible to extend the strategy learned from the training corpus with handcrafted rules that include expert knowledge or specifications about the task [25, 31].

In this chapter, we discuss the tremendous potential of Data Science to improve several aspects of Natural Language Processing and Speech Technologies research and development, including speech recognition and synthesis, natural language processing, dialog management, and natural language generation. From our point of view, as performance improves, more people will use conversational interfaces. With more usage, there will be more data that the systems can use to learn and improve. And the more they improve, the more people will want to use them. Given this cycle, it can be expected that conversational interfaces will see a large uptake for some time to come and that this uptake will be accompanied by enhanced functionalities and performance for Digital Business.

The remaining of the chapter is as follows. After this introduction, Sect. 2 is used to define conversational interfaces, describe the history and evolution of this kind of interfaces, and present the main components in their general architecture. Sections 3, 4 and 5 describe the main possibilities and benefits of the application of Data Science and Big Data to develop these components, respectively related to the natural language processing, dialog management, and natural language generation tasks. Finally, Sect. 6 presents a list of concluding remarks.

2 Defining Conversational Interfaces

Conversational interfaces can be defined as computer programs that engage the user in a dialog that aims to be similar to that between humans [11, 14, 19]. The first serious attempts to build talking systems where initiated in the XVIII and XIX centuries when the first automata that imitated human behavior were constructed. For example, Baron Von Kempelen developed in 1770 an automata that produced whole words and short phrases. In 1857, Josef Faber constructed the machine Euphonia, which imitated the mechanism of human speech production and had a 16 keys keyboard with which any word could be formed. These first machines were mechanical, it was not until the end of the XIX century when scientists concluded that speech could be produced electrically.

At the beginning of the XX century J. Q. Stewart built a machine that could generate vocalic sounds electrically and during the 30s, the first electric machines were built, for example Vocoder, an speech analyzer and synthesizer developed in Bell Laboratories that could be operated by a keyboard. At the same time, the first translating machines were patented in France and Russia, constituting one of the first systems to automatically process natural language.

During the 40s, the first computers were developed and some prominent scientists like Alan Turing pointed out their potential for applications demanding intelligence. This was an starting point that fostered the research initiatives that in the 60s yielded the first conversational agents appeared. For instance, Weizenbaum's ELIZA, which was based on keyword spotting and predefined answers. Although their behavior was considered almost human-like by some users, in practice these systems did not understand the user's inputs.

In the 70s appeared the research area of computational linguistics in which natural language processing is based, established on the theoretic work developed from the 50s by Chomsky, Montague and Wood. In the 70s also the first rule-based speech synthesizers appeared. Spoken conversational systems can be considered to appear in the 80s with the first telephony services based in DTMF, which were commercialized at the same time as the first dictation systems. The first research initiatives related to spoken conversational systems appeared in the DARPA Spoken Language Systems programme in the USA and the Esprit SUNDIAL in Europe.

In the 90s, the first corpora development and system evaluation efforts yielded big shared resources such as Wordnet. Telephony dialog systems based on isolated word recognition was a growing market and the first dedicated companies appeared. Benefiting from the continual improvements in the areas of speech recognition, natural language processing and speech synthesis.

From the beginning of the current century, there has been particular interest on data science and statistical approaches. Prospective challenges and future research directions reported by the experts in the area of spoken conversational interfaces have changed to adapt to the milestones and advances achieved during their history. For instance, at the beginning of the new century, the experts anticipated that the most important research guidelines would be related to improving success of the different components in the architecture of these interfaces (more robustness in speech

recognition, richer vocabularies and enhance algorithms for natural language understanding, less restricted dialog management models, high quality speech synthesizers with more natural voices, etc.).

Technological advances in Artificial Intelligence, language technologies and device technologies have directed these future guidelines towards the achievement of more complex goals, such as advanced reasoning, problem solving, context-awareness, portability, adaptiveness, proactiveness, affective intelligence, use of data-driven techniques, ubiquity, multimodality and multilinguality. Some of the larger scale conversational interfaces with which the general public has become familiar are those of the smartphones have currently to hundreds of millions of people around the globe.

As per the previous section, conversational have currently a relevant place in the Big Data infrastructure for various reasons. Depending on the approach used for creating an SDS, there may exist varying levels of interaction with Big Data. Usually systems require data training to provide them with a method to recognize and synthesize speech, generate natural language, or learn a strategy for dialog management. Alternatively, considering the 'three versus' of Big Data (Velocity, Variety and Volume) are all matched—a conversational system has lots of input data to respond to with little time to consider a response (velocity), often must be able to cope with auditory and textual data (variety) and often the amount of data used for training models is significantly large.

Current conversational interfaces are typically defined by their objectives, either by the class of 'goal-driven' or by 'non-goal-driven', depending on their purpose [27]. This classification refers to the purpose of the systems in questions, which in turn affect the requirements posed for the system. Goal-driven systems are designed to reach a specific goal as dictated by the user, though ultimately limited in the number of outcomes (e.g. obtaining scheduling information, making remote purchases, making reservations, etc.).

Non-goal systems are those in which there is no defined 'end-goal'. Due to the open nature of a this kind of conversational interfaces, they must be able to handle a larger set of possible user statements, requests and vocabulary. Common examples of non-goal-driven systems would be those used in mobile devices such as Apple's Siri, Microsoft's Cortana, or Amazon Echo. It is worth noting that whilst the smartphone Chatbots are well known, the value of the technology is often found in the goal-driven systems. In a survey from 2016, 80% of business leaders stated that they either "already used or planned to use Chatbots by 2020" when asked which technologies they had implemented or were planning to do so.[1]

Figure 1 shows the high-level architecture used upon which spoken conversational systems are built. This general architecture consists of three main components: natural language processing, dialog management, and natural language generation.

Note that Speech Recognition and Speech Generation components are optional, as a system that receives and outputs text is also possible, the rest of the compo-

[1]Business Insider: 80% of businesses want chatbots by 2020. http://www.businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12 Last accessed: 01/12/2017.).

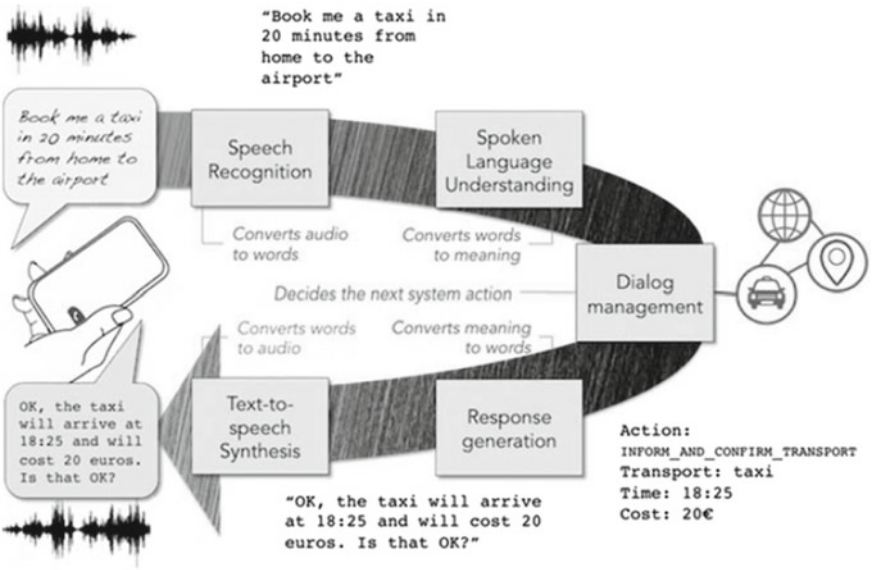


Fig. 1 Components of a spoken language conversational interface

nents are usually mandatory. Automatic Speech Recognition (ASR) is the process of obtaining the text string corresponding to an acoustic input. It is a very complex task as there is much variability in the input characteristics with can vary depending on the linguistics, the speakers, the interaction context and the transmission channel. Hidden Markov models (HMMs) have been used in speech recognition since the late 1970s and in combination with Gaussian mixture models (GMMs) have been the dominant method for modeling variability in speech as well as its temporal sequential properties. Deep Neural Networks (DNNs) have recently replaced GMM-HMM models and are now being used extensively in current ASR systems.

Within each component lie various sub-tasks that must be performed to deliver the required output. For instance, the natural language understanding task requires domain identification, intent identification and semantic parsing; dialog management requires state tracking and dialog act generation; natural language generation includes subtasks for sentence planning and surface realization, etc. The following sections describe the main connection among these modules and tasks, Data Science, Machine Learning techniques, and Big Data.

3 Natural Language Understanding

The Natural Language Understanding (NLU) component converts natural language sentences (i.e. human language) into a set of data that can be understood by the system. These data are task-related pieces of information; for instance, *What does*

the user want to do?, *Where does the user want to go? When does the user want to depart?*, etc. Ultimately, the objective is boiling down natural language to the bare minimum, scraping away the extra grammar that natural language entails to be able to conclude specifically the key details and ideas that a user wants to convey.

Whilst the task of an NLU is complex, measuring the success of an NLU component is fairly inherent to the nature of the problem. The percentage of sentences that are 'correctly understood' can be considered the accuracy of an NLU. Responsibilities of being able to piece together context from previous utterances within a conversation fall to that of the dialog manager, and would not be considered the responsibility of the NLU.

Semantic parsing is the approach to taking a natural language sentence given to a system and being able to understand to some level the meaning of the statement(s). For instance, it is with semantic parsing that a system can be programmed to understand that the following sentences each have different syntax, though semantically maintain the same meaning:

John repaired the computer within a week.
It took John a week to repair the computer.
Repairing the computer took John a week.

All provide the following information:

Agent: John
Subject: computer
Duration: A week

Various approaches to performing semantic parsing exist, falling under two main categories: compositional semantics and distributional semantics. Compositional semantics is an approach to parsing text by breaking it down into its smaller sections of clauses, phrases and words. With the knowledge of where in a sentence's structure a word falls, this knowledge is then combined with the categorization of the words according to their lexical features (i.e. is a word a noun, a verb, a preposition etc.). This serves primarily as a comparison to the distributional semantic approach and as further context for the conversational system architecture ahead of the dialog manager analysis.

Distributional semantic parsing is an approach to utilizing the words and context that surround a word to be able to understand its meaning. The approach derives from the distribution hypothesis, which states that "the degree of semantic similarity between two words (or other linguistic units) can be modeled as a function of the degree of overlap among their linguistic contexts".

In practice, this means that after enough texts have been analyzed, it might be the case that there are many sentences containing "going to Madrid","driving to London", "rowing to the shore", etc. The purpose of this example is to show that from these sentences, it could be extrapolated that any word following "to" in a sentence is a location. This in turn provides value in the conversational system when a user states "I would like to buy a train ticket to Barcelona" as the system could (if it

were to be using this form of distributional semantic parsing) deduce that Barcelona is the desired destination for the user's journey.

Domain identification and intent identification both serve the purpose of classifying and understanding about what the user is communicating (e.g., if a user asks to book a table at a Chinese restaurant, the domain here is 'restaurants' and the intent 'to make a reservation', and in this instance the slot for restaurant type would be satisfied by 'Chinese'. The reason for having domain and intent identification is to provide the system with (essentially) questions to be asked. For instance, to use the same example once more (however simple), being able to identify that the request falls under the 'restaurant' domain then supports the system in being able to identify the intent. Consider that a conversational interface could have three different 'intents' in its definition of the 'restaurant' domain (e.g., book reservation, cancel reservation and request opening hours). Restricting the number of potential 'intents' ensures that the ability of the system to identify the correct 'intent', which is then, in turn, used to better resolve the users query or problem.

Various supervised classifiers have been proposed in practice to classify both the domain and the intent present within human-machine dialog conversations. Some of the classifiers used are as follows [27]:

Support Vector Machines (SVM): SVMs, typically used for binary classification, have been extended to Multi-Class Classifiers (MCC) for the purposes of text classification. By training an individual classifier per 'class' (in our case, possible domains), 'class versus everything_else' classifiers are built, one for each desired class. Upon constructing each of these classifiers, each individual observation is classified by each individual SVM, and the classifier with the highest 'confidence' return is taken as the 'successful' classifier. As an illustrative example, for a conversational system for restaurant table management, one might choose to categorize all user queries into three buckets: 'make a reservation' (x), 'change a reservation' (y), and 'cancel a reservation' (z). To set up a (basic) MCC, there would then be constructed three classifiers, and for each classifier the chance that a given statement falls into the main class (i.e. not the 'others') is returned. For instance, passing the sentence "I would like to change my booking" may return:

- x versus others: Chance of $x = 10\%$, chance not $x = 90\%$
- y versus others: Chance of $y = 85\%$, chance not $y = 15\%$
- z versus others: Chance of $z = 40\%$, chance not $z = 60\%$

In this example, the MCC would return that the given sentence is classified as y = "change a reservation" and the conversational system could then proceed accordingly.

Maximum Entropy (ME): This approach is a method of estimating a probability distribution from given data. ME derives from the assertion that one should use the most uniform models that satisfy given constraints [17]. For instance, consider the conversational system from the previous section and its user intents (make/change/cancel a reservation), if analysis (or otherwise) informs us that 80% of statements to be received that contain "to book" can be considered as the 'make

a reservation' class. If a statement received by the system then receives a statement that does contain the phrase "to book", then the expectation is that there is an 80% chance that the statement is a "make a reservation" user intent, whereas there remains a 10% chance the intent is to change a reservation and a 10% chance the intent is to cancel a reservation (as the uniform class distribution is assumed). Of course, the above model would be easy to train due to the lack of constraints, but would not have much utility. To develop the constraints that are required for ME, word counts are used as features, and the model trained via relating the high word counts in certain texts with their corresponding classifications [18].

Convolutional Neural Networks (CNN): Typically, CNNs have been used to classify images, they can also be employed to take advantage of the structure of textual data, by (for example) taking the rows as individual letters and the columns as the correspondent with the letter in the text. Performing this transformation to the data allows for the application of Convolutional Neural Networks to be applied to sentences in natural language. Using this approach, a model has been trained by Microsoft (accessible via Azure cloud services) on the Amazon categories dataset, which contains 2.38 million sentences that have been classified over seven product categories (books, electronics, home and kitchen, clothing, movies, health, sports).

This use-case of a CNN for text classification is put forward as an example as to how this could be used in other scenarios. Identifying the category of a product can be considered as not being hugely distinct from identifying the domain of a user query or, depending on the training data, the classifier could be used to classify user intent directly. Note that this is only one approach to using CNN for text; here the approach was to train the data at the character level, but it is feasible to consider that a model could be trained, instead, at a word level, though in a model of this sort for a human-machine dialog system, the vocabulary that corresponds with the texts (e.g., the words that replace the individual letters of the alphabet) must be carefully considered.

For a supervised classification system to function, a manually labeled corpus must be given to train the model. Two important restrictions exist for the use of supervised classification techniques for domain and user intent classification. Firstly, the reliance on manual labeling of text and corpora is a bottleneck on classification, in part due to the large amount of data at hand (and that depending on the domain in which a text lies, the way the data is labeled could be different). Ultimately, manual labeling is laborious and time-consuming (and, in turn, expensive). Additionally (though this is context-dependent as this is more likely to be less of an issue for smaller systems), the approach to labeling data is not a standardized method, implying that not only may people be unfamiliar with the labeling approach (resulting in either poor quality labeling, a difficulty in interpreting results).

Unsupervised classifiers avoid both pitfalls, in that manual labeling is (intuitively) not required, and that the reliance on the manual tags is non-existent. Various unsupervised classifiers have been developed and trialled, though less work has been carried out on unsupervised techniques.

Query-likelihood Clustering (QLC): The query-likelihood approach takes the user's utterance (the input statement) as a query, and searches for similar queries in a similar manner to a search engine would. Upon receiving these similarity results, they are processed through the clustering technique. Tagging words by their function within a sentence (except 'content' words, e.g. names of 'things', which instead were stemmed so that only roots of the words persisted) is a method that was found to be successful in query-likelihood clustering, as the results implied that the significance of uttered sentences in a task-oriented domain lay in the content words more-so than the prepositions, determiners, etc. [4].

With this categorization performed, the approach was then to take each statement in the corpus and obtain the list of similar statements in the corpus to that original statement. This is then repeated for every statement in the corpus and using this overall set of data points ultimately as how the data can be clustered.

A number of toolkits incorporating AI and Machine Learning techniques are currently available to facilitate developing the described tasks for natural language processing [13, 14]: Microsoft Bot Framework,[2] Microsoft LUIS,[3] IBM Watson,[4] Amazon Alexa,[5] Google Actions,[6] Facebook Messenger Developer,[7] Api.ai,[8] Wit.ai,[9] and many others.

4 Dialog Management

Once a statement from a user has been processed by the NLU (such that the only information to be passed from the NLU to the dialog manager (DM) is the information deemed necessary to do so), the DM must select the optimal action to take. Exactly what 'optimal' means in this context depends on the purpose of the system, but often this optimization refers to being able to retrieve all of the required information from the user to be able to carry out the goal of the conversational system. The way these decisions are made will be the focus of this thesis, though for the purposes of this section this will not be touched upon here.

The types of outputs that are generated by the DM are essentially sets of instructions that tell the NLU what needs to be said to the user. For instance, if a user asks a goal-driven conversational system based on recommending restaurants, 'show me an Italian restaurant in Madrid', the hypothetical DM might understand that the desired food type is Italian, and that the location is Madrid, however the DM could be pro-

[2] https://dev.botframework.com/.

[3] https://www.luis.ai/.

[4] https://www.ibm.com/watson/.

[5] https://developer.amazon.com/alexa.

[6] https://developers.google.com/actions/.

[7] https://developers.facebook.com/products/messenger/.

[8] https://api.ai/.

[9] https://wit.ai/.

grammed to follow this kind of scenario by requesting a price range, as a result of many users in the past asking this question and extended conversation lengths. The DM can be considered the component that allows conversational interfaces to partake in conversations, as opposed to purely one phrase requests, as often these systems will be designed in a manner that permits a level of permanence for the user's responses.

In general, a dialog manager should be able to consider the history of a current conversation to be able to converse with a user. Without this history, a spoken conversational system can respond to sentences but conversations would be out of reach (for this reason, goal-oriented spoken conversational systems often require the user to state their intent before proceeding to perform actions).

A dialog strategy is needed for a dialog manager; when should the system take initiative in the conversation, how can errors in understanding (system or human) be identified and corrected, when should confirmation of understood information be checked with the user, etc. Arguably, the performance of a conversational interface is highly dependent on the performance of its dialog manager, hence the focus on the DM in spoken dialog systems, this point serving as a motivating factor for the use of the principles of data science and big data.

The most widespread methodology for machine learning of dialog strategies involves modeling humancomputer interaction as an optimization problem using Partially Observable Markov Decision Processes (POMDP) [30–32]. To understand what is meant by a POMDP one must first understand what an (standard) Markov Decision Process (MDP) is a model in which a set of 'states' exist (only one of which can be active at any one time/turn) that reflect an aspect (or some aspects) of the real world. An MDP contains five key sets of information, as follows: a set of states, a set of possible actions, a set of transition probabilities, a function to calculate rewards, a 'discount factor' that allows the model to increase or decrease the weight of possible future rewards, and an initial state.

The end goal of an MDP is to be able to define a 'policy'; i.e. a set of rules that dictate what action should be taken at any given state; note that the manner in which this policy is defined is by maximizing the possible rewards. An additional note to make here is that MDPs adhere to the Markov property: that the state and reward at the next time step in the model will depend only on the state and action taken prior [23]. The application of MDPs to model the dialog management process, the policy is obtained by optimizing the highest reward for the length of the human-machine 'conversation'.

This approach usually manages goal-oriented systems (in that the number of 'slots', i.e. the values the system seeks to fill to perform the user requests) by determining the number of states in the MDP by the number of slots the system needs to fill. The states are used to reflect the status of the slots ('confirmed,' 'known' and 'unknown') and the actions provide an avenue to change these status to 'confirmed' by way of 'request n slots', 'verify n slots', 'verify all', or 'quit.' The alternative approach to the action set is to consider the actions as the speech act (what is the goal of the next statement), a slot name (to what data will the user's answer be relevant) and a slot value (for the instances in which data confirmation is being solicited) [23].

The goal with the MDP is to find the policy that will maximize the expected value of the reward function. Different techniques exist depending on whether transition probabilities and rewards can be known in advance—where they cannot, the need is to use an algorithm that directly interacts with the DM to learn these values (described as an 'online' function, the contrary aptly-named an 'off-line' function).

One approach to the policy maximization for an online function is the use of the Q-learning algorithm [29], an approach in which quality-values are maintained for every possible combination of state and action, which reflect the value of the policy function after performing the action at that state. The optimal policy will be devised from these values, as the policy will select the action with the highest return at each state.

To take advantage of this reinforcement learning, often texts, corpora, etc. are automatically generated such that the model can be learned though this of course has drawbacks in that the data is false, and does not train a model as accurate as genuine human conversation. The advantages to using an MDP to model a dialog manager lie in the ability to plan and use this reinforcement learning to capitalize on the data. The issue with using an MDP is that the assumption that all states are 'observable' is made (i.e., that the MDP knows exactly at which state the model is at any given time in the process). The issue is that due to noisy conditions, misheard words, misunderstood sentences, etc. the need to be able to accommodate this uncertainty is a high priority. To quantify this as an issue, due to the usage of spoken conversational systems in noisy areas, such as while driving, in public spaces, etc., it could be a considerable percentage of errors in the words that are understood orally by the ASR.

To amend for this issue of uncertainty, Partially Observable Markov Decision Processes (POMDPs) are used in their stead. A POMDP assumes that an MDP exists, that the states, actions, etc. are present but that the states are not directly observable. A starting state is assumed and following states are modeled by a transitional probability. Instead of knowing the state, the model is assigning possibilities to which state the model can be in at any given time (based on previous beliefs and actions). Formally, a POMDP extends the definition of an MDP by adding an additional set of observations that can be made about the world, a set of probabilities for an observation, and an initial belief state.

A POMDP asserts that at each time-step in the model, the world (i.e. the scenario that is being portrayed) is in a given state. As this state is non-observable, a belief state, b, is distributed over the states in S, where $b(s)$ is the probability of the world being in state s. The POMDP selects the next action (a) to take at each time step based on the value of $b(s)$, receives a reward and this changes the state of the model. However this state remains non-observable, instead an observation is obtained and the belief state updated.

Using the reinforcement learning inherent to the POMDP allows the suitability of the system to increase—however dependency on human interaction (especially if this is a live system in place with end-users) entails certain risks, due to the varying expectations in place for artificial intelligence. In one instance, there have been several users that provide an inflated level of satisfaction out of a sense of 'politeness' as opposed to intrinsically and objectively answering accurately. Alternatively, as

there are many people that hold such large expectations for AI and anticipate a near-perfectly natural conversation to take place, due to this typically not being realistic for the most part, these reviews will be overly negative. On the whole, these results in the use of end-user reviews being somewhat unreliable, particularly as this type of review is likely to be considering the spoken conversational system as a whole, whereas the purpose of the information is to train the POMDP, solely part of the dialog manager [31].

Other interesting approaches for statistical DM are based on modeling the system by means n-grams models, graph-based models, or using Bayesian networks. **N-grams models** are based on predicting the next user action based on the preceding N user and system actions. However, again due to the lack of existing data upon which models may be trained, an approach is to reduce the model to the use of bigrams. The main advantage of this approach is that the model is entirely probabilistic and therefore domain-independent. However, the issue with using this model type is that the model is non-constrained with regards to the user-actions, any user action may follow the system response, disregarding the dialog history.

Due to this, dialog policies based on these models change often as the user is changing its goal. The extension to this work was the development of the Levin model [12], to allow for the generation of more realistic dialogs, primarily by restricting the estimated probabilities for feasible combinations of user response to system actions. Whilst this model achieves a more realistic dialog regarding responses, the drawback to this is the same as the general N-gram issue—that each response depends only on the system action prior, and not the rest of the dialog history (hence the contradictions that occur).

To resolve the issue of goal inconsistency generated by the N-gram approach, [24] proposed a **graph-based model** by combining a rule-based structure for situations that are goal-dependent, along with probabilistic modeling to approximate the 'randomness' in a human user's behavior. Here, a 'path' that the dialog can follow is mapped out as a network, via 'choice points' and actions. Some of the choice points are solely based on reaching the goal of the dialog ('deterministic') or, alternatively, random choices based on previous data that do not affect the final goal ('probabilistic'). Whilst this approach ensures that trained models follow simulated users with consistent goals, the bottleneck is that the models are highly domain-dependent, and additionally, having to map the dialog flow for each system reduces adaptability, and in more complex cases would become nigh-on unfeasible.

The use of models based on **Bayesian networks** is focused on avoiding the effort required to construct a model such as that seen in the graph-based models [16]. The goal of the user is recorded as a table of slot name, slot values and associated status (known/unknown/etc.). These status, combined with the number of times pieces of information have been provided during the dialog is recorded by the model, are used to generate a priority for the user to provide this information.

The statistical methodology proposed in [8] is based on modeling the dialog management process by means of a classification process. The goal of the DM is then to select the best possible system action, after considering the history of the dialog prior to the current time-step. The concern raised is that due to the potentially

large amount of data that could be held in a dialog, the number of different dialog histories could be very long in practical dialog systems. To solve this problem, a 'dialog register' (DR) is introduced: a manner in which the entire history of the conversation can be recorded at each time-step.

The dialog register records the status of which of the variables that the system is aiming to obtain from the user during the dialog (empty, filled with a high confidence value, filled with a a low confidence value). The DR and the last system response are considered as the input feature set for the classifier and this provides the probability of selecting each one of the possible system responses to continue the dialog. Different classification functions can be employed for the practical implementation of this approach.

5 Natural Language Generation

As context for this section, and as previously stated in the introduction section, the output of the DM is 'what' the system should say (in a non-linguistic form, i.e. semantic information), and the goal of the natural language generation (NLG) component is to convert this information into one or more sentences in natural language for a human user. For instance, the NLG can take the output from the DM 'ask-departure-time' and formulate this into natural language ('at what time would you like to depart?').

It is important to note that measuring the success of an NLG component in any given system is difficult due to there being more subjectivity in the goal of the component. Compared to an NLU (which can be assessed via a quantitative analysis of the number of words/utterances that were understood out of the total given), or the DM (for which various qualitative assessments exist,for instance, how quickly did the user achieve their goal), the goal of the NLG is less discrete. First and foremost, the goal of an NLG is to convey the information that it has been passed from the DM to a human user in a way that they are able to fully comprehend the given information. However, a user's satisfaction with a conversational system depends on other factors. For instance, if the system responses to a user are not entirely grammatically accurate, it may be more difficult for a user to understand than intended.

Whilst it can be considered dependent on the conversational system in question (what is the purpose, who is the audience, etc.), the tone (read: style) of the language returned to the user can be considered important. If the conversational system is used in a context aimed primarily at children, this should be reflected in the vocabulary used. If the conversational system is aimed at retaining customer attention for a significant period of time, the language used should not push the user away, the language used could aim to avoid boring the user and maintaining their engagement, etc.

Ultimately, because of the various purposes and contexts within which a conversational system may be employed (as the field a conversational system looks to cover gets larger, the more likely the need will be to implement a NLG system that expands

beyond simply using look-up tables and templates to return a response, though this may be sufficient in certain cases; for instance one NLG component could be as simple as a look-up table for dialog manager outputs (e.g. in a table reservation system, one can picture the responses of 'how can I help you', 'at what time would you like your table', 'your reservation has been canceled', 'okay' and 'thank you, goodbye' as suitable for the vast majority of possible queries based on a make/change/cancel reservation system). Equally, template responses could be used in a similar approach ('your train leaves at [departure time]').

As stated, the bottleneck for these approaches is that the generation of these sentences for any system with a relatively large set of responses would be expensive (as each would need to be labeled). In smaller goal-oriented conversational systems this is less of an issue, though for non-goal-oriented systems this approach is either expensive or borderline impossible (once phrases need to be returned with many variables, the risk of high volume of potential responses is opened due to the permutability of said variables). Note that where speech is the way to communicate with the conversational interface, a text-to-speech synthesizer is required to be able to communicate back to the user via the same medium. To deliver a 'how to say' from the 'what to say' delivered by the DM, an NLG uses both 'sentence planning' and 'surface realization' [27].

According to [22], three main sub-tasks are required for building a sentence planning natural language generation system. Lexicalization and referring expression generation aim to choose the words that can be used to describe the desired concepts and entities (this is where the output of the DM is transformed). It is localization that provides the words to describe a concept/relationship, whilst referring expression generation is the task in which words are selected to describe an entity. The lexicalization process searches for a word to refer to the concepts in the data, whilst the referring expression generation searches for words to refer to the entities. Running the data through these processes could provide several sentences.

6 Main Conclusions

Conversational interfaces allow people to communicate with their devices in a more intuitive and natural way. Current conversational interfaces have been made possible by a renaissance in Artificial Intelligence, specially thanks to the dramatic improvements brought by deep learning techniques and powerful processors that support their computations in the fields of automatic speech recognition, spoken language understanding, and dialog management. The integration of these technologies and algorithms on small devices such as smartphones have also made possible an impressive increment of the number of potential users of these devices, number of new application domains and more complex interaction scenarios.

These developments are also linked to advances in technologies related to the Semantic Web, which enable an almost instantaneous access to large amounts of unstructured and structured data repositories on the Internet to take advantage of Big

Data main principles and apply them for the construction of enhanced methodologies and models for natural language processing, automatic speech recognition and synthesis, and dialog management.

As a result of these technological advances, user acceptance of technologies such as the conversational interface has increased, leading to increased adoption and consequently producing more data from which systems can learn, in turn resulting in further improvements in the technology. In this chapter, we have explored the main technologies and components that make up a conversational interface and the main alternatives to apply Data Science to develop these components. As it has been described, there are great opportunities for making conversational interfaces become a new revolution in digital business, but also many challenges ahead related to areas of particular interest, such us how to provide a satisfying conversational experience and how to evaluate this kind of interfaces.

References

1. Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C., Morgan, N., et al. (2009). Developments and directions in speech recognition and understanding. *IEEE Signal Processing Magazine*, *26*(3), 75–80.
2. Bohus, D., Grau, S., Huggins-Daines, D., Keri, V., Krishna, G., Kumar, R., et al. (2007) Conquest—an Open-Source Dialog System for Conferences. In *Proceedings of of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, USA (pp. 9–12).
3. Cavazza, M., de la Cámara, & R.S., Turunen. (2010). How Was Your Day? a Companion ECA. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, Toronto, Canada (pp. 1629–1630).
4. Ezen-Can, A., & Boyer, K. (2013). Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of 6th International Conference on Educational Data Mining* (pp. 2–9).
5. Fryer, L., & Carpenter, R. (2006). Bots as Language Learning Tools. *Language Learning and Technology*, *10*(3), 8–14.
6. García-Márquez, F., & Lev, B. (2017). *Big data management*. Springer International Publishing.
7. Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., et al. (1995). Multilingual spoken-language understanding in the MIT Voyager system. *Speech Communication*, *17*, 1–18.
8. Griol, D., Callejas, Z., López-Cózar, R., & Riccardi, G. (2014). A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer Speech & Language*, *28*(3), 743–768.
9. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., & Jaitly, N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *82*, 82–97.
10. Hodson, H. (2014). The first family robot. *New Scientist*, *223*(2978), 21–22.
11. Lee, G., Kim, H., Jeong, M., & Kim, J. (2015). *Natural language dialog systems and intelligent assistants*. Springer.
12. Levin, E., Pieraccini, R., & Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 11–23.
13. McTear, M. F. (2017). Future and emerging trends in language technology. In *Second International Workshop on Machine Learning and Big Data, FETLT 2016, Chap. The Rise of the*

Conversational Interface: A New Kid on the Block? (pp. 38–49). Springer International Publishing.
14. McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface: Talking to smart devices*. Springer.
15. Melin, H., Sandell, A., & Ihse, M. (2001). CTT-bank: A speech controlled telephone banking system—An initial evaluation. In *TMH Quarterly Progress and Status Report (TMH-QPSR)* (Vol. 1, pp. 1–27).
16. Meng, H. H., Wai, C., & Pieraccini, R. (2003). The use of belief networks for mixed-initiative dialog modeling. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 757–773.
17. Nigam, K., Lafferty, J., & Mccallum, A. (1999). Using maximum entropy for text classification. In *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering* (pp. 61–67).
18. Nigam, K., McCalum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2/3), 103–134. https://citeseer.nj.nec.com/nigam99text.html.
19. Ota, R., & Kimura, M. (2014). Proposal of open-ended dialog system based on topic maps. *Procedia Technology*, *17*, 122–129.
20. Pieraccini, R., & Rabiner, L. (2012). *The voice in the machine: Building computers that understand speech*. The MIT Press.
21. Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, *16*, 171–194.
22. Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Journal of Natural Language Engineering*, *3*(1), 57–87.
23. Schatzmann, J., Weilhammer, K., Stuttle, M., & Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, *21*(2), 97–126.
24. Scheffler, K., & Young, S. 2001. Corpus-based dialogue simulation for automatic strategy learning and evaluation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001). Workshop on Adaptation in Dialogue Systems*, Pittsburgh, USA (pp. 64–70).
25. Suendermann, D., & Pieraccini, R. (2012). One year of contender: What have we learned about assessing and tuning industrial spoken dialog systems? In *Proceedings of SD-CTD'12* (pp. 45–48).
26. Vaquero, C., Saz, O., Lleida, E., Marcos, J., & Canalís, C. (2006). VOCALIZA: An application for computer-aided speech therapy in Spanish language. In *Proceedings of IV Jornadas en Tecnología del Habla*, Zaragoza, Spain (pp. 321–326).
27. Wang, X., & Yuan, C. (2016). Recent advances on human-computer dialogues. *CAAI Transactions on Intelligence Technology*, *1*(4), 303–312.
28. Wang, Y., Wang, W., & Huang, C. (2007). Enhanced semantic question answering system for e-learning environment. In *Proceedings of of 21st Conference on Advanced Information Networking and Applications (AINAW'07)*, Niagara Falls, Canada (pp. 1023–1028).
29. Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292.
30. Williams, J., & Young, S. (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, *21*(2), 393–422.
31. Young, S., Gasic, M., Thomson, B., & Williams, J. (2013). POMDP-based statistical spoken dialogue systems: A review. *Proceedings of IEEE*, *101*(5), 1160–1179.
32. Young, S., Schatzmann, J., Weilhammer, K., & Ye, H.: The hidden information state approach to dialogue management. In *Proceedinds of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Haway, USA (Vol. 4, pp. 149–152).
33. Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., et al. (2000). JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 85–96.

After 2017: Managers Exit and Banks Arise

Takafumi Mizuno

1 Introduction

A drastic change will occur in Japan. A private bank, Bank of Tokyo-Mitsubishi UFJ, will issue an original currency which is referred to as MUFG coin in fall 2017 [1, 2]. The news shocked us with two reasons: (1) a private bank violates the right of issuance of the central bank and the seigniorage of the country, and (2) this is a cryptocurrency guaranteed by one trillion dollars (100 trillion yen) bank deposit and its infrastructure consists of blockchain technologies. Blockchain infrastructures have been realized in actual economic world. By using the infrastructure, we can move worth and credit without human resources beyond borders of countries. Worth and credit are guaranteed by computing resources.

Combing such currencies and smart contracts changes economic environments. Concept of smart contract is proposed by Szabo in 1997 [3], and now realized on the blockchain infrastructures. Smart contract is business contracts described as programs in if-then rules. If conditions of the rules will be satisfied, then the transactions contract will be executed automatically.

These technologies, blockchain and smart contract, change jobs of managers of companies and banks. By the changes, I describe how changed the economic models. Managers are described in linear programming of operations research, and banks are represented as an economic agent.

1.1 Blockchain Technology and Smart Contract

Blockchain is sequence of blocks $B_i = (y_{i-1}, n_i, h(X_i), X_i)$. The function h is a hash function. The value of the function is easy to calculate, but the inverse of the function hard to calculate. X_i is data, and y_{i-1} is hash value of previous block. y_{i-1} is calculated as follows

T. Mizuno (✉)
Meijo University, 4-102-9, Yada-Minami, Higashi-ku, Nagoya, Aichi, Japan
e-mail: tmizuno@meijo-u.ac.jp

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_4

$$y_{i-1} = h(y_{i-2} \mathbin{+\!\!+} n_{i-1} \mathbin{+\!\!+} h(X_{i-1}) \mathbin{+\!\!+} X_{i-1}), \tag{1}$$

where $s \mathbin{+\!\!+} t$ is concatenation of bit expression of s and t. n_i is a sequence of random bits which is referred to as *nonce*, and it holds

$$\text{a value starts with sequence of zero} = h(y_{i-1} \mathbin{+\!\!+} n_i). \tag{2}$$

To find the nonce, large computing resources are needed. Thousands of computers connected in Internet try to find the nonce, and the computer that find the nonce update blockchain. In Bitcoin protocol, some fees are paid to the computer [4]. Why blockchain cannot be falsified depends on difficulty of finding the nonce. An attacker which try to falsify data of a block has to find nonces which output same hash values in all blocks.

In cryptocurrencies realized by blockchain, data of each block of blockchain is some bundle of transaction histories. Records of transactions in blockchain increase monotonically. The monotonicity, the difficulty of falsification, and openness of transaction histories enable us move economic values on Internet.

We can see the blockchain as distributed database. Smart contract is an application on the database. Contracts of business are described on the database, and when conditions of the contracts are satisfied then the contracts will be executed. In cases of contracts with paying fees, cryptocurrencies are paid. Histories of execution of contracts and histories of paying fees are also recorded in blockchain.

Smart contracts need not any servers. Peers, which are computers joining to find nonces, can realize decentralized autonomous organization (DAO).

2 Models of Managements

DAO separates ownership of corporations and managements of corporations. A main mission of managers is executing contracts given by owners. Managers also must submit business reports without falsifications. If jobs of managers are described as program in blockchain, then the contracts can be executed automatically, and blockchain technologies provide validity of reports of states of corporations without falsifications. A problem that we have to consider is how to describe jobs of managers.

Kinoshita [5, 6] has been proposed abstract models of managements of corporations and governments. The models are represented in linear programming of operations research.

Kinoshita claims there are two strategies in corporations' managements: management in Thetical phase and management in Antithetical phase. Management in Thetical phase is strategy of corporations when economy in prosperity, and management in Antithetical phase is strategy of when economy in recession.

When economy in Thetical phase, corporations maximize their business profits. The maximization is represented in a primal problem of linear programming that is formulated in

$$\max \sum_{j=1}^{n} p_j q_j, \tag{3}$$

$s.t.$

$$\sum_{j=1}^{n} c_{ij} q_j \leq d_i, \quad i = 1, 2, \ldots, m, \tag{4}$$

$$q_j \geq 0, \quad j = 1, 2, \ldots, n. \tag{5}$$

Each variable means as follows.

q_j How many units of the product j are produced.
p_j Profits obtained by producing one unit of the product j.
c_{ij} Costs in the accounting subject i to produce one unit of the product j.
d_i How many debts the corporation can rent for the account subject i.
m How many accounting subjects there are.
n How many kinds of products are produced.

This is a maximization problem of the corporation's profits represented in $\sum_{j=1}^{n} p_j q_j$.

When economy in Antithetical phase, corporations minimize their debts. The minimization is represented in a dual problem that is represented in

$$\min \sum_{i=1}^{m} u_i d_i, \tag{6}$$

$s.t.$

$$\sum_{i=1}^{m} u_i c_{ij} \geq p_j, \quad j = 1, 2, \ldots, n, \tag{7}$$

$$u_i \geq 0, \quad i = 1, 2, \ldots, m, \tag{8}$$

where

u_i Unpaid balance rate of the accounting subject i; $u_i = 1 -$ (amortization_rate).

This is a minimization problem of the corporation's debt represented in $\sum_{i=1}^{m} u_i d_i$.

Owners of corporations decides select which optimizations will do, then the parameters in the equations are detected automatically by solving the optimization problem. Implementing the models on the blockchain infrastructures can replace jobs of managers to computations of software.

3 Banks as Economic Agent

In macroeconomics, there are three economic agents: households, firms, and governments. Each agent does not represent particular organization. A function or an aspect of people in nation is referred to as economic agent. Since banks are particular

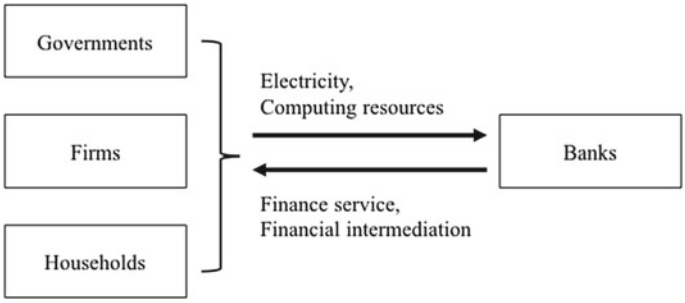


Fig. 1 Governments, firms, and households are economic agents in macroeconomics. Banks, which is new economic agent added in this article, provide finance service and intermediation to other economic agents, and other agents provide electricity and computing resources to banks

firms, banks are not represented in economic agent in spite of their important functions. Banks provide two functions: finance, and intermediation. By using blockchain infrastructures, the functions of banks become functions provided by people; banks become an economic agent.

Banks as economic agent provides finance services, financial intermediation. Other agents provide electricity and computing resources to the banks (Fig. 1).

4 Discussions

Kinoshita [5, 6] also provides management models of government. There are two managements: management in Thetical phase, and management in Antithetical phase. So, we can represent managements of government in program of smart contract.

When economy in Thetical phase, governmental actions do not need to expand the economic scale. A required action of the government is described in

$$\min \sum_{j=1}^{N} F_j L_j, \tag{9}$$

$$\sum_{j=1}^{N} S_{ij} L_j \geq U_i, \quad i = 1, 2, \ldots, M, \tag{10}$$

$$L_i \geq 0, \quad j = 1, 2, \ldots, N. \tag{11}$$

Each variable means as follows.

L_j Rate of national loans remain for the administrative service j; increasing the rate increases expenses of the service j.

F_j Demand for funds as national loans for the administrative service j.

S_{ij} How many increments of satisfaction of the resident i by increasing one unit for cost for the service j are.
U_j Required level of total services for the resident i.
M The number of residents.
N How many services are provided by the government.

This is a minimization of national loans; the government is required its finance reform.

When economy in the Antithetical phase, the government has to control demand side to balance the supply and demand. A required action of the government is described in

$$\max \sum_{i=1}^{M} Y_i U_i, \tag{12}$$

$$\sum_{i=1}^{M} Y_i S_{ij} \leq F_j, \quad j = 1, 2, \ldots, N, \tag{13}$$

$$Y_i \geq 0, \quad i = 1, 2, \ldots, M, \tag{14}$$

where

Y_i The amount of public money to increase satisfaction by one unit for the resident i.

This is a maximization of residents' satisfaction. The formulation means that the government is required fiscal stimulus. It is generally referred to as fiscal policy (or financial policy).

Like corporations' management, policies of governments can be described in program on smart contract. If government or central bank issues cryptocurrency on blockchain, then the system can be realized; people select which policy have to do, and the selected policy will execute automatically.

In Kinoshita's management model of corporations, d_i and u_i are important parameter decided by banks and regulations of governments. These parameters affect which economy goes prosperity or recession. By using smart contract to decide the parameters, the parameters can be decided with excluding biases of particular administrators and can be executed fair.

Initial coin offering (ICO) is an application on blockchain to collect resources of a project. The project issues the original cryptocurrency, and people which support the project buy the currency. If the project successes, the value of the currency increases.

The ICO is a direct finance. Usual banks lend the money to projects. The limit of the lending is some times larger than owned capital of the banks. The amount larger than owned capital is gain by credit creation. If direct finances is realized, the credit creation is unnecessary.

On blockchain infrastructures, a new kind of corporation will arise. The corporation treats service products whose marginal costs are zero. The products of the

corporation is provided by executing smart contract. Its costs, which are only electricity and computing resources, are provided by households or other corporations.

Corporations providing services on Internet, such as Google, realize low marginal costs of products because they do not produce goods, and it is one of reasons why they have become gigantic companies. But they are paying costs of electricity of their servers and costs for cooling them; marginal costs cannot be zero. New corporations on blockchain infrastructures replace their costs to accounting subjects of other corporations or consumers. Marginal costs correspond to c_{ij} in (4). If the values are zero, corporations can realize infinite profits theoretically.

5 Conclusions

I described how economic models are changed by blockchain infrastructures. Managements of corporations are whether maximize profits or minimize debt, and it represented in solving optimization problems. Blockchain technologies can automatically fulfil business contracts and achieve managements. I implied that managers become name of software. Paradoxically, I discovered why corporations need managers, that is, information around corporations is not monotone and is easy to falsify.

And I introduced a new economic agent "banks." In traditional macroeconomic model, there are three economic agents: households, firms, and governments. The banks, which are aspects and functions of people of nation, provide finance service and financial intermediation to other agents, and other agents provide electricity and computing resources to banks.

In this article, I supposed that blockchain infrastructures are provided by every economic agent. There are two types of blockchain: public blockchain, and private blockchain. I treated public blockchain in this article. Computing resources of private blockchain are provided by particular organization. Some central banks and governments start developing systems of cryptocurrency on blockchain infrastructures. They are private blockchain. Analysis of how private blockchain changes economic models is a future work.

References

1. BTMU testing virtual currency, considering launch for customer use, The Japan Times News, 10 Jun 2016. Retrieved October 31, 2017, from https://www.japantimes.co.jp/news/2016/06/10/business/btmu-testing-virtual-currency-considering-launch-customer-use/#.WfhE-0yKVBw.
2. Rizzo, P. (2016). Japans Largest Bank Latest to Develop Own Digital Currency. Retrieved October 31, 2017, from https://www.coindesk.com/japans-bank-tokyo-mitsubishi-mufg-coin/.
3. Szabo, N. (1997). The Idea of Smart Contracts. Retrieved October 31, 2017, from http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_idea.html.

4. Nakamoto, S. (2009). Bitcoin: A Peer-to-Peer Electronic Cash System. Retrieved October 31, 2017, from https://bitcoin.org/bitcoin.pdf.
5. Kinoshita, E. (2015). Thetical and antithetical business management. *Journal of Business and Economics*, *6*(6), 1086–1096.
6. Kinoshita, E. (2015). A proposal of thetical economy and antithetical economy by using operations research techniques. *European Scientific Journal, July 2015 edition, 11*(19), 29–48.

Synergies Between Association Rules and Collaborative Filtering in Recommender System: An Application to Auto Industry

Liming Yao, Zhongwen Xu, Xiaoyang Zhou and Benjamin Lev

1 Introduction

With the rapid development of internet technology, information technology and computer science, we have entered globally into the era of big data since 2010. Data shows that in 2014, the world generated the quantity of 23 TB data every day, which is 920 times than that in 2012. Just as IDC predicts, in the next few years, data volume will increase in 50%, where globally data volume will arrive at 40 ZB in 2020, and China's data volume will be 8.6 ZB, accounting for 21% of all countries in the world. As we know, big data has become the most important strategic resource for enterprises and society, which leads to many studies and applications. Even though, we are easily confused by the exponential data inevitably. Hence, it is vital for us to apply suitable methods to make full use of these data.

As for applications, many recommender systems are carried out in recent years to relieve the problem of overload information, as they has access to suggest the right items (products or services) to particular users through mining their explicit and implicit preferences with diverse information filtering technologies. Up to now, applications of recommender systems involve various domains including movies, music, television programs, books, documents, websites, conferences, tourism scenic spots, learning materials and so on. For example, YouTube developed a personalized video recommendation algorithm, which recommends suitable videos for consumers through mining consumers' preference among historic records. Amazon applies collaborative filtering (CF) to recommend the most similar items to consumers according to items they bought early. Google News apply the most suitable news to different people with the method of data mining. Meanwhile, scholars have been continually

L. Yao · Z. Xu
Business School, Sichuan University, Chengdu 610065, China

X. Zhou (✉)
International Business School, Shaanxi Normal University, Xi'an 710062, China
e-mail: x.y.zhou@foxmail.com

B. Lev
LeBow College of Business, Drexel University, Philadelphia 19104, USA

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_5

researching recommender system. Many methods have been applied to recommender system, such as content-based, collaborative filtering-based and hybrid methods.

Above all, precision recommendation has made some achievements, but researches in the auto industry is very few. In China, demand for cars has been slowly down, many small size companies have difficulty in surviving. On the one hand, these companies have little access to benefit from three leaders of data because of restriction of their scale and capitals. On the other hand, without the support of intelligent decision, small-size companies are hardly able to provide suitable and personalized cars for consumers. Hence, precision recommendation, known as a low-cost and high production approach, is suitable for solving the problems confronted by small-size automobile companied.

In a word, we aim to propose a novel recommender system, specifically applied to auto industry for two problems. One is to discover the target customers and find commonness among them. The other is to recommend suitable goods to them. Different from general retail commodities, there is some main difficulties in auto industry. First, cars are updated frequently, but a customer owed as usual one car for a time. Second, consumers don't buy cars frequently. The above difficulties result in a lack of rating information on products from customers so traditional recommendation algorithm cannot be simply applied to automobile industry. In the last section of this paper, an empirical analysis is presented to verify novel model. The main contribution of this paper is as follows:

(1) Our approach provides a theoretical framework for studying the recommender system problem that combining some methods together.

(2) In this paper, we develop a synergistic recommendation process including association rules and collaborative filtering, where we summarize a framework for better applying CF method.

(3) In empirical study, item-attributes rating matrix based on customers' dynamic browsing preference takes place of traditional rating matrix, which can relieve the problem of data scarce.

The remainder of this paper is structured as follows. In Sect. 2, the recommendation techniques are reviewed and analyzed. Section 3 presents our novel recommender system and Sect. 4 is a empirical study on auto industry. Section 5 is conclusion and future works suggested.

2 Literature Review

In the previews articles, main recommendation techniques include some traditional methods, such as content-based methods and collaborative filtering-based methods.

2.1 Collaborative Filtering (CF) Method

Collaborative filtering (CF) method is commonly used in recommender system, which is based on explicit rating feedback on the items/products by like-minded users (known as neighbors) [12]. Collaborative filtering method began with memory-based CF method, which can be divided into user-based and item-based CF method. User-based CF method, exploit the most similar customers to a user and generate recommendations accordingly [13]. Item-based CF method recommend items that are similar to those the user has loved/purchased in the past [18]. By contrast, user-based CF method tends to provide hot items, while item-based CF method is able to recommend personalized items. However, such methods has limits because users can be neighbors only if they have rated common items, and the fact is there are few or no common items they rated. Subsequently, hybrid CF method came into being.

Unfortunately, scalability and scarcity problem are two main reasons for poor recommendation of CF methods. In traditional CF models, rating matrix is based on users' feedback, however, in fact, users often forget to submit rating information, which leads to the scarcity of rating matrix. In addition, Cold start problem is specific expression of data scarcity, that's to mean, a system cannot generate personalized and relevant recommendations for a user who has just registered into the system. To cope with above problems, many methods have been studied. Cantador apply the idea of Cross-domain to build better rating matrix by finding rating information about the user's preferences in a related source domain, which consequently improve prediction performance [4]. However, it is hard to define the similarity between the target domain and the cross-domains. Implicit data can indicate users' preferences by providing more evidences and information through observations made on users' behaviors. Seo et al. [25] proposed a friendship strength-based personalized recommender system that recommends topics or interests users might have in order to analyze big social data, using Twitter in particular. Qiu transformed features of users, items, words into the same vector space, and completed rating matrix where every entry is a rating that is a measure of how much a user likes an item according to users' interested aspects and items' highlighted polarities on aspects [23]. Lepri et al. [16] analyzed social network structure to obtain users' preference. Farnadi et al. [7] obtained users' preference information through social media. To overcome the problem of data sparsity, other methods are shown as follows, such as, dimensionality reduction methods [2, 6, 10, 14, 24, 32], neutral network [22, 25, 29, 31], slope one [15, 28] and so on. Among the different real-word domains, there is not a unique solution for them, and we should choose the appropriate methods.

2.2 Content-Based (CB) Method

Content-based (CB) recommendation techniques recommend commodities that are similar to items previously preferred by a specific user [10]. Different from CF

method, the core of CB recommendation techniques is text mining, including machine learning, data mining and neutral network etc., which means that there is no cold-start problem. Text mining discovers special key words or labels to describe a consumer through analyzing his (her) attributes, such as gender, age, explicit description of interests, etc., and then build users' interest profiles based on these key words or label, later, comparing products' characteristics to interest profiles to process recommendation. Decision trees are usually ternary, the answer gained from one user is either Like, Dislike, or Unknown. Per her answer, some levels come into being, and each tree node represents a group of users [21]. However, results vary from different tree nodes, which, to some extents, leads to complexity of computation. Association rules have access to output interesting relationships among a large set of data based by describing the relationships between the typical past users' navigation paths. For example, in daily life, A–B means people who like product A like product B [20]. Association rules are able to generate rules whose support and confidence are higher than the defined min-support and min confidence. However, too many data sets lead easily to redundancy of rules, which accordingly decease the accuracy of recommendation. Thus association rules are usually combined with other methods [17, 20, 27]. Rough set, on the premise of analyzing and dealing with the imprecise, inconsistent and incomplete information, can be applied for attribute reduction or feature selection [5, 8, 30]. Neutral network derives from artificial neutral network (ANN), and evolves into deep learning. ANN is known as connectionism, parallel distributed processing, neuron-computing, natural intelligent systems and machine learning algorithms. The basic advantage of ANN is that they can learn from representative examples without providing special programming modules to simulate special pattern in the data set [1, 9, 19]. To explore more abstract situation, improved deep learning algorithm method comes into being. Artificial intelligence, image recognition, etc. are gradually used in recommender system, such as Alphago developed by Google.

However, content-based method generally suffer from the problems of limited content analysis and over-specialization [26]. For example, personal information is hard to be obtained for privacy issues. For movie recommendation domain, a movie of the same genre or having the same actors may be recommended to a user according to movies already seen by this user. Because of this, the system may fail to recommend items that are different but still interesting to the user.

3 The Framework of Recommender System

In this section, we propose a two stage recommender system to reduce marketing cost at the meantime satisfy more users' demands. Based on big data, we go deep into consumers' characteristic, preference and behavior, then adopt association rules and collaborative filtering methods to find potential users and recommend suitable goods.

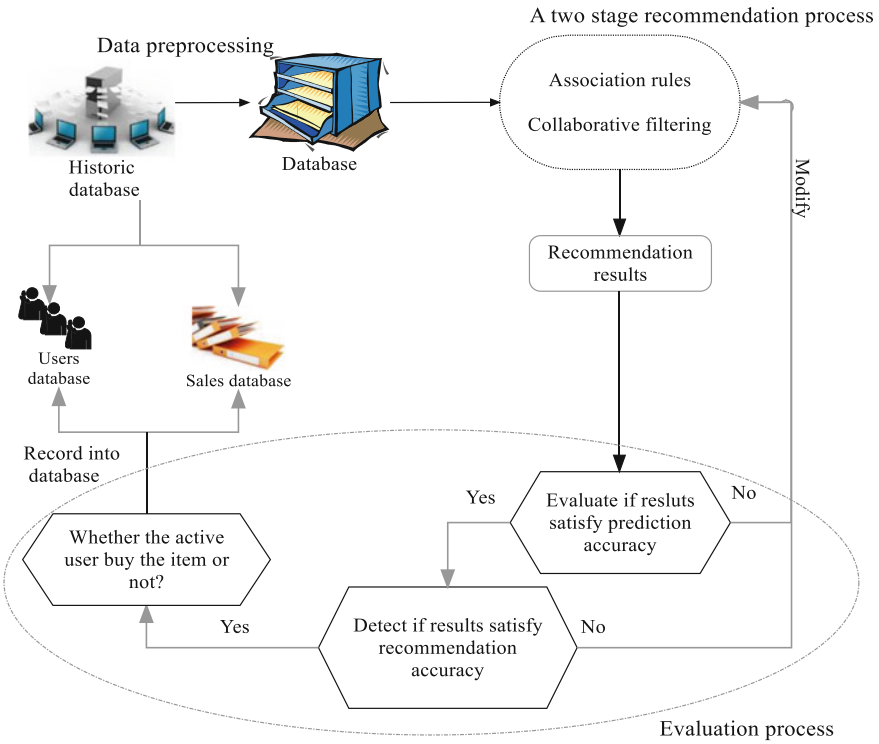


Fig. 1 Recommender system framework

3.1 General Introduction to the Proposed Recommender System

In this section, we introduce the proposed recommender system on the whole, which consists of three parts, data pre-processing, recommendation and evaluation, as shown in Fig. 1 and Table 1. In the next three section, we plan to introduce explicitly by steps. Table 1 describe the main methods and contents of the recommender system.

3.2 Data Pre-processing

Data preprocessing is helpful to understand and complete the task of recommender system. It includes data cleaning, integration, simplification and other steps, and it can provide useful data set for recommender system by reducing dimensions of the original data. There are three explicit processes: first, deep learning method is

Table 1 Methods description

Components	Description
Step 1: data pre-processing	Data cleansing, integration, simplification and other steps
Fuzzy C-mean	Discretization
Rough set theory	Attribution reduction
Deep learning	Filling the missing data
Output:	Users' matrix, items matrix and user-attribute matrix
Step 2: recommendation	Discover aimed users and calculate similarity
Association rules	Discover aimed users
Collaborative filtering	Predict their preference
Outputs:	Result
Step 3: **evaluation**	
MAE and RMSE	The accuracy of prediction
Precision and recall	The accuracy of recommendation
Outputs:	Feedback

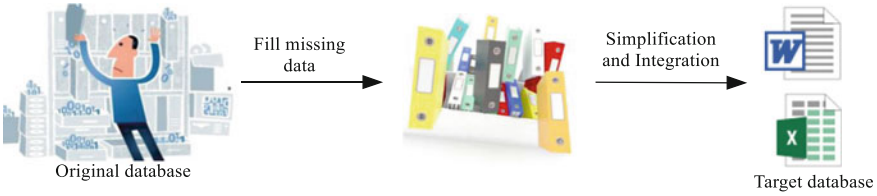


Fig. 2 The flowchart of pre-processing

applied to fill the missing data. namely, clustering algorithm is applied to cluster the incomplete data sets. Based on the clustering results, we use automatic encoding model to discover the characteristic of incomplete data sets, and then output complete data sets [3]. Second, semantic ambiguity should be solved. Data is collected from different channel, which increase easily computing complexity and difficulty. For example, a same attribute from different channel may be named differently. Hence, the main purpose of this step is to unitize data sets. Third, we simplify date sets relevant closely to the task of recommender system. In this paper, rough set theory is applied to simplify dimensionality of attributes. It is worth pointing that rough set theory can only solve discrete data. Then we must transform the other data types to Booleans. Category data can be directly mapped to new Boolean data, for example, gender can be divide into male (S1) and female (S2), then the value equal to 1 means according attribute. As for numeric data, fuzzy C-mean classification is adopted. Such as age (Fig. 2).

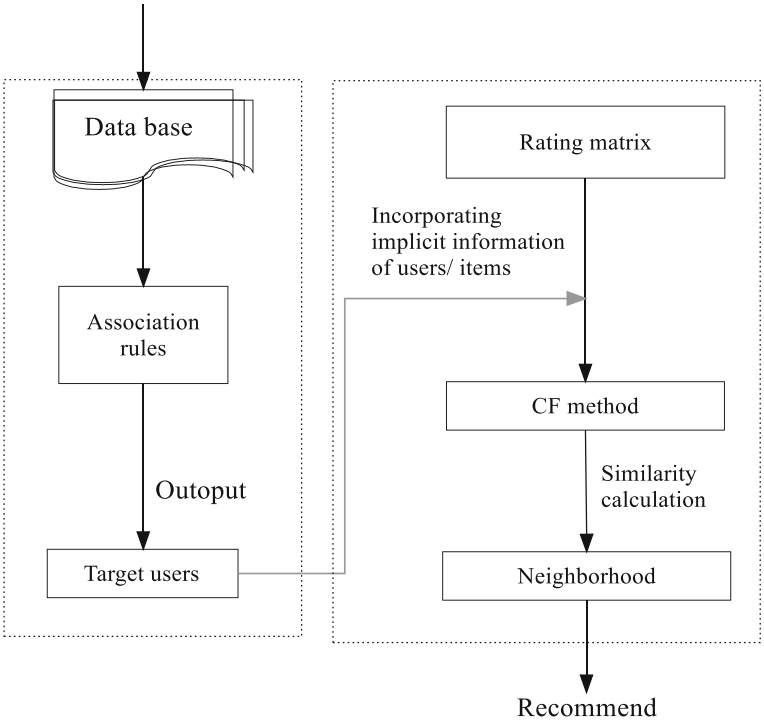


Fig. 3 The two stage flowchart of recommendation

3.3 Recommendation Process

As shown in Fig. 3 recommendation consists of two stages: the first stage use association rules as a means of finding potential customers; in the second stage, CF methods are applied to realize recommendation.

3.3.1 Costumers Identification based on Association rules

The purpose of association rules is to describe the database before predicting and recommending. We apply this method to classify customers and discover relationship between users' characteristic and buying behavior.

As we know, Personality influences people' decisions. People with similar personality traits are likely to have similar tastes, while different characteristics lead to different preference and buying behavior. For example, female likes romantic movies while male likes crime movies. It's common to see people in the same position have a similar perspective, and like the same type of things as usual. Even more,

individuals with similar incomes have the similar consumption level. Therefore, using large scale demographic data is appropriate in recommendation process.

In this section, before carrying out association rules, we have to build a decision table at first, then based on the implicit correlation matrix, we can output some rules that satisfying some defined restrict.

A pseudo code is shown to explicitly describe the association rules algorithm.

In all, we can divide users into different groups based on rules, a group of users who has purchase invention are paid more attention.

Its's worthy indicating that, certain indices, including support, confidence, and lift, can be used to validate the quality of the rules derived using association rules. The support of the rule Support ($X \longrightarrow Y$) means the probability that a transaction contains X and Y. Confidence ($X \longrightarrow Y$) means the conditional probability that a transaction having X also contains Y. Lift measures the change ratio between the probability that a transaction having X also contains Y and the probability that a transaction having Y. (1) if the value of lift is equal to 1, it means there is no correlation between X and Y; (2) if the value of lift is less than 1, it means the rules is invalid; (3) if the value of lift is more than 1, it means the rule we obtain is of value.

3.3.2 Recommendation Based on CF Method

Not enough information about ratings of the users might attribute to low recommendation accuracy. However, in real-word life, most cells of the user-item rating matrix are empty because of explicit feedback is not always available. In such cases, discovering similar users or items (neighborhood formation) becomes a challenge. So in this subsection, we integrate a framework for applying CF method, shown in Fig. 4, which concludes three types of CF method, which uses user-item matrix, item-attribute matrix and user-item matrix. The first is based on users' rating records, and the following are based on user/item implicit information. Namely, each user/item is represented by an N-dimensional attributes which carry the rating information for each item.

Similarity calculation is needed before obtaining the top-N nearest neighbors. Equations 1 and 2 describe the similarity between user u and v, and between item p and q between item p and q respectively [11].

$$Sim\,(u, v) = \frac{\sum_{p \in I(u) \cap I(v)} \left(r_{u,p} - \bar{r}_u\right)\left(r_{v,p} - \bar{r}_v\right)}{\sqrt{\sum_{p \in I(u) \cap I(v)} \left(r_{u,p} - \bar{r}_u\right)^2} \sqrt{\sum_{p \in I(u) \cap I(v)} \left(r_{v,p} - \bar{r}_v\right)^2}} \tag{1}$$

$$Sim\,(p, q) = \frac{\sum_{u \in U(p) \cap U(q)} \left(r_{u,p} - \bar{r}_u\right)\left(r_{u,q} - \bar{r}_q\right)}{\sqrt{\sum_{u \in U(p) \cap U(q)} \left(r_{u,p} - \bar{r}_p\right)^2} \sqrt{\sum_{u \in I(p) \cap I(q)} \left(r_{u,q} - \bar{r}_q\right)^2}} \tag{2}$$

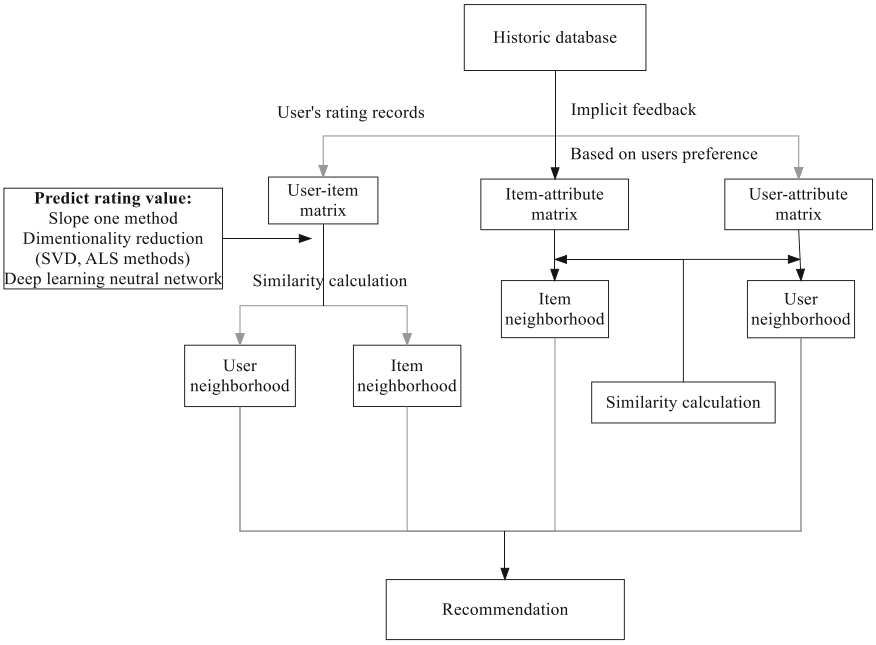


Fig. 4 A integrated collaborative filtering framework

3.4 Evaluation

The main evaluation aspects include prediction accuracy and recommendation accuracy. First, we put recommendation into effect if the prediction result is satisfactory. Second, we get feedback from consumers, if consumers are satisfactory with our recommendation, we keep this recorder into sales database, otherwise, modify the recommender system. We emphatically introduce some indexes, consisting of MAE, RMSE, Precision and Recall, as shown in Eqs. 3–6 [11].

$$MAE = \frac{\sum_{i=1}^{n} \left| r_{ui} - \hat{r}_{ui} \right|}{n} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(r_{ui} - \hat{r}_{ui} \right)^2}{n}} \tag{4}$$

where r_{ui} and $\hat{r}_{ui}$ is real rating score and predicted rating score by user u to item i, n represents the numbers to be predicted. The smaller MAE or RMSE is, the more accurate the recommender system.

Table 2 Methods description

	Like	Not like
Recommend	Recommend, like (a)	Recommend, no like (b, bad result)
Not recommend	No recommend, like (c, bad result)	No recommend, no like (d)

$$Precision = \frac{\sum a}{\sum a + \sum b} \tag{5}$$

$$Recall = \frac{\sum a}{\sum a + \sum c}. \tag{6}$$

The more the values of precision and recall, the more accurate the recommender system (Table 2).

4 Application of Auto Industry

4.1 Problem Description

Cost-effectiveness ratio is one of the most failures that may lead to serious sequences in many industries. Among the industries whose inputs for a single user are high, auto industry has been kept with aggressive advertising input. iResearch points that auto industry invests the most money in advertisement, as shown in Fig. 5. However, according to statistics from China Association of automobile manufacturers in 2015, China produced 24.5033 million vehicles for the year, which had an increase of

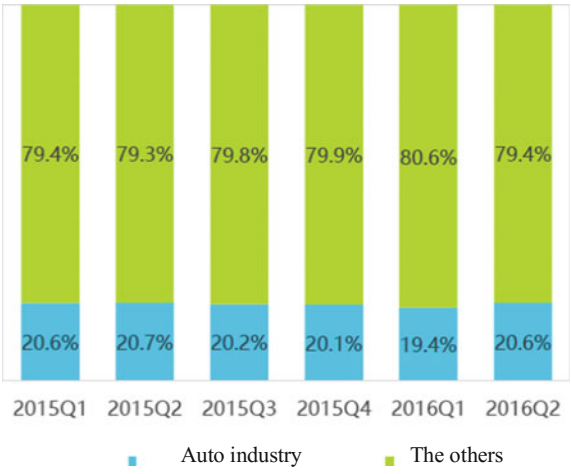


Fig. 5 Ratio of advertising investment

3.25%, but sold 24.5976 million vehicles, which had an increase of 4.68%. But the increase ratio is lower than that in 2014. In other words, in 2015, China's vehicle market shows modest growth trend, slower than the last year.

Generally speaking, the demand of car in Chinese markets has slowly down, even worse, large-size companies put pressure to small size companies. On the on hand, small size companies have little access to benefit from three leaders of data because of the restriction of scale and capitals. Without the support of intelligent decision, small size companies are hardly able to provide suitable and personalized cars for consumers. With the time goes by, small-size companies is hard to survival. On the other hand, small size companies often sponsor some promotional activities to attract customers, but faced with some "wrong" users, these activities don't have any significant effect. Above all, too much money has been put into advertising with unsatisfactory revenue, thus, it is urgent to propose a recommender system suitable for small size companies.

Meanwhile, there is some main difficulties in auto industry. First, cars are updated frequently, but a customer owes as usual one car for a time. Second, consumers don't buy cars frequently and are unable to tell about their preferences on all available items and cannot rate millions of them. The above difficulties result in a lack of rating information on products from customers, the similarity between two users or items cannot be calculated since there is not enough information about ratings of the users and thus, the recommendation accuracy becomes very low. Therefore, traditional recommendation algorithm cannot be simply applied to auto industry.

4.2 Deployment

Above all, we obtain users' characteristics and behavior information of customers from data company, and then recommend top-k cars to the active customer. Data (180 records totally about customers' searching behavior) we use concludes characteristics (Age, Educational background, Marital status, Monthly income, Vocation, Having a car or not), behavior information (Searching behavior on auto websites, The amount of login, the time of browsing automobile website one day, the duration staying in automobile website one time) and company's sale information (want to buy car or not). After data preprocessing, we apply association rules to judge a group of people who has more possibility to buy cars, namely, discovering target customers.

As users' requirements of cars' configuration are concrete and constant for a time, and an active user might not buy cars frequently. Then, CF method applied in this section is to exploit user preference reflecting 2-dimensional internal attributes ("have browsed and have not browsed") to generate recommendations.

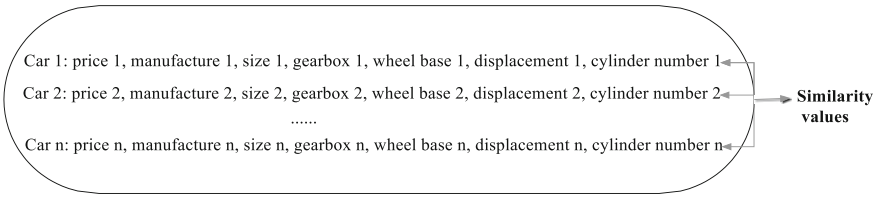


Fig. 6 The similarity values between each two cars

Step 1: Based on costumers' dynamic browsing records, we build item-attribute matrix.

Step 2: Calculate similarity and find the top-k closest neighbor sets, as shown in Fig. 6. In this step, Pearson Correlation Coefficient and Cosine Vector Similarity methods are applied. Then we represent the top-3 closest items for the active user considering the number of sample size.

4.3 Result Analysis

4.3.1 Identifying Target Customers

When running association rules algorithm, we set up a low support (1%) and a high confidence (50%) to discover the significant correlation among users' characteristics and behavior with his/her purchase intention. In Table 3, we present 9 rules obtained.

Take Rule 5, for example, represents that "If someone is female, AND age is between 21 and 26, AND educational background is undergraduate, AND she is a single person, AND she works as a state-owned company, AND she likes to search car series, AND she has no car yet, AND she browse auto websites one time a day, AND each time she browsing auto websites lasts ten minutes to sixty minutes, THEN she want to buy a car."

Except for the obvious rules obtained, we can get some implied information by making comparative analysis. First, comparing rule 1 to rule 2, the common characteristics change in the time staying in car websites of each login, which implies that more users staying in car websites for 10–60 min of each login are likely to buy a car. Second, when rule 3 is compared to rule 4 and 8, we know a user (no matter male or female) with a higher salary has more possibility to buy a car, which conforms to real phenomenon. Third, when considering searching behavior on auto websites, we find a user making a price comparison is more likely to buy a car according to rules 5 and 6.

Table 3 Result based on association rules

Rules		Support (%)	Confidence (%)	Lift
1	Female, 21–26, undergraduate, single person, 2000–6000, price comparison, have no car yet, one time, less than 10 min $\Longrightarrow$ want to buy	2.21	57.14	1.246
2	Female, 21–26, undergraduate, single person, 2000–6000, price comparison, have no car yet, one time, less than 10–60 min $\Longrightarrow$ want to buy	3.31	66.67	1.454
3	Female, 21–26, undergraduate, single person, 2000–6000, the others, have no car yet, one time, less than 10 min $\Longrightarrow$ want to buy	2.21	66.67	1.231
4	Female, 21–26, undergraduate, single person, less than 2000, the others, have no car yet, one time, less than 10 min $\Longrightarrow$ don't want to buy	9.73	93.33	1.724
5	Female, 21–26, undergraduate, single person, 2000–6000, state-owned staff, search car series, have no car yet, one time, 10–60 min $\Longrightarrow$ want to buy	1.11	66.67	1.454
6	Female, 21–26, undergraduate, single person, 2000–6000, institutions, price comparison, have no car yet, one time, less than 10 min $\Longrightarrow$ want to buy	1.66	75	1.636
7	Female, 21–26, undergraduate, single person, less than 2000, state-owned staff, search car series, have no car yet, one time, less than 10 min $\Longrightarrow$ don't want to buy	1.11	66.67	1.231
8	Male, 21–26, undergraduate, single person, less than 2000, the others, price comparison, have no car yet, one time, less than 10 min $\Longrightarrow$ don't want to buy	1.1	100	1.847
9	Male, 27–40, undergraduate, single person, 2000–6000, stated-owned staff, price comparison, have no car yet, one time, less than 10 min $\Longrightarrow$ want to buy	1.66	66.67	1.454

4.3.2 Recommending Cars Based on CF Method

In this section, we apply CF method to run a database containing 50 sets browsing records, data chosen in this subsection within a group of person female, 21–26, undergraduate, single person, 2000–6000, price comparison, have no car yet, one time, less than 10–60 min. In Table 4, we choose the top 3 similar cars to recommend. Take the first result for example, if a customer browsed a car (such as Honda, XR-

V2015–1.5LLXiCVT) in website, we know his/her preference for cars' attributes and we can recommend the most 3 similar cars to the customer, namely, Guangzhou Honda, YARiS L ZHIXUAN–1.3E CVT; PEUGEUT 4008 (export)–2.0LCVT; DONGFENG FENGSHEN–AX31.4T.

If we choose another group of person female, 21–26, undergraduate, single person, 2000–6000, price comparison, have no car yet, one time, less than 10 min, the numbers of recommending cars for these group of customers may be different. In a word, the number of recommended items is variable faced with different groups of customers.

5 Conclusion and Future Work

In this paper, we provide a novel recommender system, which contains data preprocessing, recommendation and evaluation. The recommendation process is based on an integrated algorithm, with the sense that the integrated recommendation can save time and ensure better performances. For the first stage, we employ association rules to divide customers into different groups and discover target customers who has more possibility to buy a item. In the second stage, we apply CF method to distinguish top-N items for recommending. In this way, it saves time cost instead of recommending to those users who have no buying intention. In empirical study, we apply the integrated recommender system to auto industry aiming to provide a low cost, large profit model to discover target customers and recommend suitable cars. Instead of user-item rating matrix, we explore implicit of items (users' dynamic browsing records) with data mining technique to build item-attribute matrix to calculate similarity and recommend the top-k items. In conclusion, the main contribution of this proposed system is as follows: (1) We build a complete framework of recommender system from data-preprocessing to results evaluation. (2) To reduce the overhead, we propose an integrated two stage recommender process combining association rules with CF method together. First, we can find customers who have more possibility to buy a car based on association rules. Second, based on the group customers chosen by association rules, we implement recommendation based on collaborative filtering. (3) We summarize three types of rating matrix to calculate similarity when using CF methods based on explicit feedback and implicit feedback respectively. As future work, it is suggested to continue the experimental evaluations in other fields like music, books and TV programs, and the explore the optimal number of k based on MAE, RMSE, Precision and Recall. What's more, we intend to explore more methods to overcome the problem of data sparsity, such as using rating matrix of similar domains and define different weights according to different domains.

References

1. Al-Alayah, W. M., Kadhum, A. A. H., Jahim, J. M., El-Shafie, A., & Kalil, M. S. (2014). Erratum to: Neural network nonlinear modeling for hydrogen production using anaerobic fermentation. *Neural Computing & Applications*, *24*, 1229–1229.
2. Barragans-Martnez, A. B., Costa-Montenegro, E., Burguillo, J. C., Rey-Lopez, M., Mikic-Fonte, F. A., & Peleteiro, A. (2010). A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. *Information Sciences*, *180*, 4290–4311.
3. Bu, F., Chen, Z., Zhang, Q., & Yang, L. T. (2016). Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud. *Journal of Supercomputing*, *5*, 1–14.
4. Cantador, I., & Cremonesi, P. (2014). Tutorial on cross-domain recommender systems. In *ACM Conference on Recommender Systems* (pp. 401–402).
5. Chen, L. F., & Chihtsung, T. (2016). Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, *53*, 197–206.
6. Chowdhury, N., & Cai, X. (2016). Nonparametric Bayesian probabilistic latent factor model for group recommender systems. Springer International Publishing.
7. Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.F., & Cock, M.D. Computational personality recognition in social media.
8. Gao, R., Tang, L., & Wu, J. (2011). A novel recommender system based on fuzzy set and rough set theory, *3*, 100–109.
9. Ghasab, M. A. J., Khamis, S., Mohammad, F., & Fariman, H. J. (2015). Feature decision-making ant colony optimization system for an automated recognition of plant species. *Expert Systems with Applications*, *42*, 2361–2370.
10. Guo, G., Zhang, J., & Thalmann, D. (2012). A simple but effective method to incorporate trusted neighbors in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 114–125).
11. Hu, Y. C. (2014). Recommendation using neighborhood methods with preference-relation-based similarity. *Information Sciences*, *284*, 18–30.
12. Huang, S. Y. (2015). Non-invasive magnetic resonance imaging (mri)—Based electrical property mapping for human tissues. In *IEEE Mtt-S International Microwave Workshop Series on Rf and Wireless Technologies for Biomedical and Healthcare Applications* (pp. 1–1).
13. Kasap, O. Y., & Tunga, M. A. (2017). A polynomial modeling based algorithm in top-n recommendation. *Expert Systems with Applications*, *79*, 313–321.
14. Kumar, B., Srivastava, A., & Kumar, P. (2016). Cosine based latent factor model for precision oriented recommendation. *International Journal of Advanced Computer Science & Applications, 7*.
15. Lemire, D., & Maclachlan, A. (2007). Slope one predictors for online rating-based collaborative filtering. *Computer Science*, 21–23.
16. Lepri, B., Staiano, J., Shmueli, E., Pianesi, F., & Pentland, A. (2016). The role of personality in shaping social networks and mediating behavioral change. *User Modeling and User-Adapted Interaction*, *26*, 1–33.
17. Li, Z., Li, L., Yan, K., & Zhang, C. (2016). Automatic image annotation using fuzzy association rules and decision tree. *Multimedia Systems*, 1–12.
18. Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments. *Decision Support Systems*, *74*, 12–32.
19. Masthoff, J. (2003). Modeling the multiple people that are me. In *International Conference on User Modeling* (pp. 258–262).
20. Najafabadi, M. K., Mahrin, M. N., Chuprat, S., & Sarkan, H. M. (2017). Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. *Computers in Human Behavior*, *67*, 113–128.

21. Nanopoulos, A., Nanopoulos, A., & Schmidt-Thieme, L. (2015). A supervised active learning framework for recommender systems based on decision trees. Kluwer Academic Publishers.
22. Paradarami, T. K., Bastian, N. D., & Wightman, J. L. (2017). A hybrid recommender system using artificial neural networks. *Expert Systems with Applications*, *83*, 300–313.
23. Qiu, L., Gao, S., Cheng, W., & Guo, J. (2016). Aspect-based latent factor model by integrating ratings and reviews for recommender system. *Knowledge-Based Systems*, *110*, 233–243.
24. Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *International Conference on World Wide Web* (pp. 285–295).
25. Seo, Y. D., Kim, Y. G., Lee, E., & Baik, D. K. (2017). Personalized recommender system based on friendship strength in social network services. *Expert Systems with Applications*, *69*, 135–148.
26. Shardanand, U. (1995). Social information filtering: Algorithms for automating "word of mouth". In *Sigchi Conference on Human Factors in Computing Systems* (pp. 210–217).
27. Sharma, D. (2016). Application of association rules in clinical data mining: A case study for identifying adverse drug reactions. *Value in Health*, *19*, A101–A101.
28. Tian, S., & Ou, L. (2016). An improved slope one algorithm combining knn method weighted by user similarity. In *International Conference on Web-Age Information Management* (pp. 88–98).
29. Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*, *69*, 29–39.
30. Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, *37*, 7838–7843.
31. Zhang, F., & Chang, H. (2006). Employing bp neural networks to alleviate the sparsity issue in collaborative filtering recommendation algorithms. *Journal of Computer Research & Development*, *43*, 667.
32. Zhou, X., He, J., Huang, G., & Zhang, Y. (2015). Svd-based incremental approaches for recommender systems. *Journal of Computer & System Sciences*, *81*, 717–733.

Information Security Research Challenges in the Process of Digitizing Business: A Review Based on the Information Security Model of IBM

Jason X. S. Wu and Shan Liu

1 Introduction

The rise of big data and cloud computing has brought all types of data analysis techniques into the process of business decision and further reshaped the business process. Analysis of various data, such as macroeconomic, enterprise operational, and consumer behavior data, has greatly improved business decisions. However, some security concerns arise in the process of digitizing businesses.

Existing security research has two categories of issues, namely, data security and system security [44]. With regard to data security, common previous topics have mainly focused on developing all types of encrypting technology to help protect security online, such as key cryptography, secure socket layer, and cookies [44]. However, industry and research communities have added "personnel security" as the third category. Unconscious and malicious organizational insiders should be responsible for nearly half of security breaches because of failure to comply with information security policies (ISPs) of firms. Organizational insiders refer to individuals who are authorized to access organizational internal information system (IS) or other assets. Security managers and researchers have considered three main types of measures to compel these insiders to follow ISPs, including security education, training, and awareness (SETA) programs, fear appeal, and system monitoring [11, 16, 22, 23, 29], which have become new research interests that emerged from the digitization of businesses. Meanwhile, in system security protection, all types of traditional system security protection tools or software have been used against endless attacks, such as firewalls, proxy servers, and virtual private networking [44]. However, the Ponemon Institute reported that 80% of businesses cannot properly manage external cyberattacks, although they spend an average of $3.5 million per year for all types of system security protection deployment [15]. Meanwhile, the

J. X. S. Wu · S. Liu (✉)
School of Management, Xi'an Jiaotong University, Xi'an 710049, China
e-mail: shan.l.china@gmail.com

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_6

recent studies and industry reports have claimed that consumer awareness of threats caused by privacy risk [8] has been intensified; moreover, consumers have become more concerned about whether organizations are sufficiently capable and willing to protect their ISec although data gatherers tell consumers what they have collected and promise that they will protect their data from illegal usage. Hence, the contradiction between the low efficiency of organizational system security protection and the increasing concerns for user privacy has emerged and has even been intensified with the increased digitization of businesses.

Additional issues, such as anonymization and data masking, lack of legal protections, patents and copyrights concerns, have emerged. Even discrimination may be enhanced by using big data analytics in the process of digital businesses. Different from the existing discrimination concerns, big data analysis allows a type of "automated" discrimination. For example, the racial or sexual orientation of an applicant is not allowed to be disclosed to the financial organization in the existing process of business decision. However, this information can be easily inferred through data analysis based on various data collected online. Thus, conducting targeted studies to explore the possible countermeasures is urgently needed.

To sum up, when we focus on the security of an organization, aside from data, IS (or IT, adopted and deployed by the organization), human behavior (including insiders and consumers of the organization), business process as the new type of object must be emphasized and its interaction with other objects (e.g., data, human and system) must be explored. Before adopting and utilizing big data analytics, organizations should consider and be requested not to infringe on consumer privacy and to avoid creating additional hidden security concerns. As discussed above, all types of ISec research and new challenges have emerged. However, knowledge about whether the academic research has responded to the requirements of industry well, especially in the process of digital business, is insufficient. Hence, a summary of existing research can serve as a guideline for industry application. It can also serve as a theoretical basis and literature in determining possible ways of managing new security challenges and provide new directions for subsequent researchers.

This chapter aims to review and compare academic studies and existing requirements of industry to provide insight into the matching extent of literature and practice. This review also further identifies some research directions, especially new emergent research topics or challenges due to the digitization of businesses for further ISec research by IS researchers. We adopt the ISec framework of IBM as a representative of the requirement of industry to cluster the existing security studies published in the mainstream IS journals in the past four years.

The following section will introduce the research method. This section specifies the manner in which journals and papers are selected and each paper is coded, from which we provide a comprehensive overview of ISec research. This section also imparts some important insights. The next section briefly presents current ISec studies based on the IBM security model, which provide us a basis to determine the four important security objects (i.e., data, human behavior, IT/IS, and business processes) in organizations. The subsequent section will further examine the interactions between different objects and some recommendations for the research and

industry communities are identified and provided. Finally, we provide a comprehensive conclusion and limitations of our review.

2 Review Method

Guidelines of Webster and Watson [66] are adopted to conduct this review through four steps. First, we decide on the research fields, search items, and criteria for the inclusion or exclusion of a paper. Second, we limit the sample after searching through all the identified sources. Third, analysis of the texts of the selected set of studies is conducted, considering the IBM ISec capability reference model. Finally, we categorize and structure the content of our review. Before introducing the details of our review method, we first introduce the IBM ISec capability reference model.

2.1 Core ISec Themes: IBM ISec Capability Reference Model

IBM always stands at the frontline of the ISec battlefield and provides the timely response and effective solutions to the security threat of industry to maintain leadership of the security market and be a competitive ISec service provider. Thus, security solutions of IBM is a good representative of the requirements of industry and can provide a good guideline for the academic community. The IBM Information Security Capability Reference Model is a comprehensive model that addresses technical, behavioral, and managerial issues related to ISec. Thus, the model supports our initial argument, which emphasizes on the four objects (i.e., data, human behavior, IT/IS, and business processes) in organizations. Using this model, IBM can help businesses with security troubles in assessing their enterprise security posture and then provide all types of measures for improving their security level. The eight security themes of this framework are shown in Table 1. Although the themes of the model cover broad areas of ISec, the assessment factors help us to limit potential research areas associated with each theme. Zafar and Clark used this model to review the security literature in mainstream IS journals from their founding to 2007 and proposed ISec economics as a ninth theme into this model [70]. This chapter will use the final model adapted by Zafar and Clark to classify the latest security research published in six leading IS journals in the past four years.

2.2 Journal Selection and Paper Identification

Journals in the IS field are selected because we focus on what IS researchers have contributed to ISec research. To ensure the quality of selected papers over quantity,

Table 1 Zafar and Clark's adaption of IBM Information Security Capability Reference Model

Security themes	Assessment	
Governance	• Strategy and information • Security policy • Security compliance	• Security risk management • Governance structure • Information security advisory
Privacy	• Policy, practices and controls • Privacy and information Management strategy	• Data, rules, and objects
Threat mitigation	• Network segmentation and boundary protection • Vulnerability management	• Content checking • Incident management
Transaction and data integrity	• Business process transaction security • Database security	• Message protection • Secure storage • Systems integrity
Identity and access management	• Identity proofing • Access control	• Identity lifecycle management
Application security	• Systems development life cycle	• Application development • Environment
Physical security	• Site management physical	• Asset management
Personnel security	• Workforce security	
Information security economics	• Information security investment	• Consumer choice

we selected six mainstream IS journals from the "Basket of Senior IS Scholars" that are deemed high quality [36, 37]. The six journals are as follows.

MIS Quarterly
Information Systems Research
Journal of Management Information Systems
Journal of the Association for Information Systems
European Journal of Information Systems
Information Systems Journals

To be included in our review, each journal article must include security or privacy as a key construct and be relevant to organization security in response to our focus. In our search for literature, which involved identifying papers on ISec in web of science core collection using keywords "security" and "privacy," we found 65 articles related to ISec in the target journals and the defined years. During the first review, we removed six articles that were not organization security studies in the actual sense but only contained references on security concept. Finally, we obtained 59 articles for in-depth coding.

2.3 Coding Methods

According to the traditional coding method, we coded each paper's "author–time," "research questions", "theoretical basis", "research method", "research findings and practical implication", and "limitation". Aside from these items, we also coded each paper as one or more themes of IBM security model suited to our purposes. One paper may relate to two or more topics. Accordingly, each paper was assessed to focus on one or more types of objects (i.e., data, human behavior, IS/IT, and business processes) and their interaction. The details of coding result are shown in the appendix.

3 Overview of ISec Research

3.1 Papers Distributions by Journal and Period

Table 2 shows that 59 ISec papers published in the selected journals from January 2014 to 2017 have focused on the organizational security. Except for 2017, published ISec papers are increasing with years, which shows the importance of ISec research in the IS field. When journals are considered, JMIS and MISQ are the top two journals that publish most ISec studies with emphasis on organizational ISec.

3.2 Contribution of ISec Research to Industry Requirements

Figure 1 shows the number of papers for each theme of the IBM ISec model. Most work of IS research community refers to the "governance" theme, which involves the development of strategic and compliance programs, mechanism, and structure. In addition, "personnel security" and "threat mitigation" are the second and third themes explored. The former mainly considers how to confine and normalize the behaviors of the insiders and users of organizations to avoid information leakage, whereas the latter focuses on threat or vulnerability detection and incidence management.

In view of the focus of this review, we are interested in the security concerns and findings in the process of digitizing businesses. This topic is closer to the "transaction and data integrity" and "privacy," which lack extensive research. We further counted the number of papers of important and less important themes in each year and observed the papers' number change with years, as shown in Fig. 2.

From Fig. 2, although themes including "governance" and "personnel security" have more relevant papers, papers in both themes decreased with years. Increasing unimportant theme includes "identity and access management". Meanwhile, "privacy", "threat mitigation", and "information security economics" seem to be increasingly important themes because papers for these themes increased with years (we

Table 2 The number of papers published by each journal in each year

Target journal	2014	2015	2016	2017	Total
Information systems research	2	4	4	1	11
MIS quarterly	2	5	2	3	12
Journal of management information systems	3	4	8	2	17
European journal of information systems	4	1	4	2	11
Journal of the association for information systems	1	1	2	0	4
Information systems journals	2	2	0	0	4
Total	14	17	20	8	59

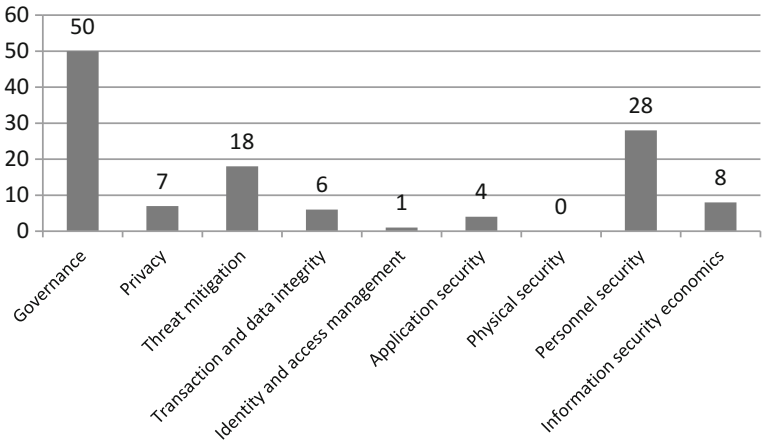


Fig. 1 The number of papers located in each theme

only searched papers from 2014 to 2017). "Application and physical security themes" seem to become farther from the sight of the IS scholars, which may be attributed to the themes being closer to computer science research topics.

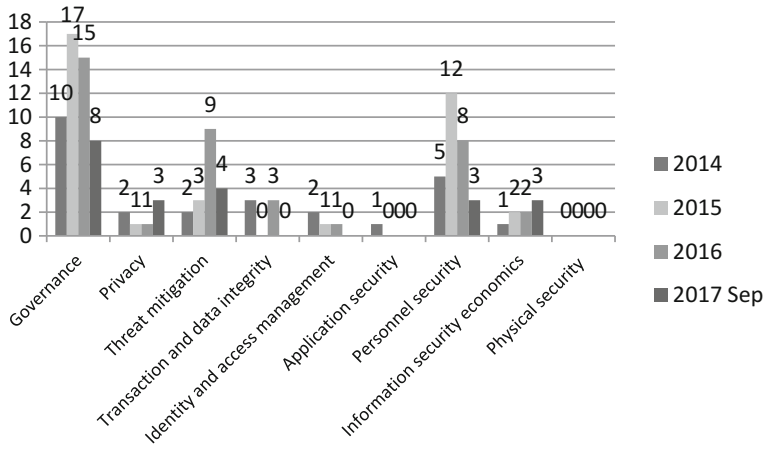


Fig. 2 The number of papers located in each theme in each year

Finally, "transaction and data integrity" seems to have no rule to identify its importance. However, when we consider the three months left in 2017, we expected more related papers would be published. Nevertheless, we can still consider the theme as an important topic because more than 10% of papers are about this topic.

To sum up, the IS community has contributed more to the following three themes of IBM ISec model: "governance", "threat mitigation", and "personnel security". Among these themes, "threat mitigation" has become increasingly important. However, the other two seem to be excessive studied by ISec scholars. "Privacy" and "transaction and data integrity" are also worth of emphasis and further analysis.

3.3 Main Theories and Methods Conducted in Each Theme

We also summarized a brief description of the research methods and theories for each theme. Table 3 shows how varied research in ISec is and how it can be advanced further. More sociological theories (especially in the "personnel security" theme) and organizational theories were adopted, and qualitative and quantitative research methods were used.

4 Research Streams Summary

In this section, we provide a brief overview of each of the articles according to theme and method of assessment.

Table 3 Theories and methods adopted in each theme

Themes	Theory basis	Methods used
Governance	Persuasion theory; Motivation theory; Social control theory; Information foraging theory; Deterrence theory; Compliance theory; Theory of knowledge retention; Accountability theory; Social network analysis; Habituation Theory; Institutional theory; Strain theory	Design science; Behavioral experiment; Modeling; Case study; Scenario-based survey;
Privacy	Theory of unintended consequences; Self-control theory; Theory of planned behavior; Social learning theory; Mindfulness theory; Universal philosophical theories of ethics; Opportunity theory of crime; Institutional anomie theory; Routine activity theory; Actor-network theory; structuration theory; Contextualism; Selective organizational rule violations model; Dual-task interference theory; Justice Theory; Sanction Theory; context-updating theory; Organizational control theory; Reactance theory; Fairness theory; Extended parallel process model; Behavioral decision-making; Traditional monitoring methods; Expectation confirmation theory; Technology Threat Avoidance Theory (TTAT)	A growth mixture model approach; Functional magnetic resonance imaging; Grounded theory Ethnography; Theory development; Approach development; Interview; Experiment
Threat mitigation	K-anonymity framework; regression tree; Privacy impact assessment; Selective organizational rule violations model; Strain Theory; General deterrence; Expectation confirmation theory; Theory of Justice; Psychological Contract theory; Institutional Theory	Algorithm; Experiment; Approach development; Theory development; Grounded theory; Scenario-based survey; A growth mixture model approach; Case
Transaction and data integrity	Persuasion theory; Motivation theory; Diffusion of innovation; Context sensitive theory; Coping theory; Interpersonal deception theory; Social distance theory; Media richness theory; Opportunity theory of crime, Institutional anomie theory; Institutional theory; Trait theory	Experiment; Modeling; Survey; Data mining/machine learning; Case study; Integration-analytics technologies; Design science; Algorithm

(continued)

Table 3 (continued)

Themes	Theory basis	Methods used
Identity and access management	Protection motivation theory(PMT); Detection tool impact theory; Context-updating theory; Justice Theory; Sanction Theory; Extended parallel process model; Behavioral decision-making; Text mining technologies; Mindfulness theory; Expectation confirmation theory; Psychological contract theory; Habituation Theory	Scenario-based survey
Application security	K-anonymity framework; Regression tree; Cultural dimensions; Migration theory; Technology Acceptance Model (TAM); TTAT; PMT; Four-factor theory; Leakage theory; Verbal cues; Static Dynamic Linguistic; Competence model of fraud detection; Information manipulation theory; Criteria-based content analysis; Scientific content analysis; Reality monitoring; Channel expansion theory; Interpersonal deception theory; Interaction adaptation theory; Text mining technologies	Functional magnetic resonance imaging
Physical security	PMT; TTAT; Economic theory TAM	Experiment; Case study; Data mining/machine learning; Survey; Algorithm
Personnel	Theory of planned behavior; Expected utility, deterrence, and ethical work climate theory	Experiment; Factorial Survey; Modeling; Survey

4.1 Governance

Organizational ISec governance aims to form a stable management framework, which includes the mechanisms, processes, and structures by which organizational ISec is controlled and directed. The governance mechanisms are realized by the development of and compliance with ISec strategies and policies. Governance processes mainly include information objective setting and pursuit in the context of social, regulatory, and market environments. Organizations should focus on security risk management. Governance structures and principles aim to build bodies that will identify the responsibility of different participants in monitoring and governing organizational security. Following is a brief discussion of how governance, as applied to ISec, has been addressed in our selected journals. Most of papers are located in the development and compliance of strategy and information security policy.

In terms of strategy and information security policy, D'Arcy et al. [14] explored how complicated and unclear ISec requirements cause "security-related stress" to employees. Bhattacherjee and Park [7] explained the reason for users to move from client-centric computing to cloud computing. Choudhary and Zhang [13] explored the impact of a change in the distribution of defect-related costs on a vendor's release time and patching strategy under SaaS. Tsohou et al. [57] proposed a framework to guide designing and implementing ISec awareness programs by considering changes that happened in an organization. Vance et al. [60] found that users' perceived accountability could be increased by the UI design of broad-access systems and further reduce their intentions to violate access policies. Steinbart et al. [55] focused on what influences users' (dis)continuance to adopt security behavior from the UI design perspective. Johnston et al. [29] explored the effectiveness of an enhanced fear appeal rhetorical framework using a hypothetical scenario research design involving three unique threat/behavior pairs that were typical of fear appeal implementations in practice. Hsu et al. [22] clarified and examined the role of extra-role behaviors and social controls in organization on ISP effectiveness. Chen and Zahedi [11] investigated the differences in security behaviors between the people from United States and China on a relatively large scale based on context-sensitive theory. Kim et al. [31] examined the effect of cultural difference on security concerns on e-transactions. Ji et al. [27] analyzed a size-based security monitoring policy with and without profiling. Choi et al. [12] developed a model that shows how the recovery measures of firms influence customer behavior online after data breach. Goode et al. [17] explored how a breached organization could best determine the optimal level of customer compensation in response to data breach. Jensen et al. [26] developed a new way to conduct security training given that some employees are used to training based on rules. Wang et al. [62] suggested that companies should improve employees' coping adaptiveness, which is combined by task-focused coping, emotion-focused coping, and avoidance coping in the process of phishing email detection. Niemimaa and Niemimaa [45] explored the manner in which the best practice of IT service provider of information security system can be converted into contextualized practices. Khansa et al. [30] investigated the cyberloafing behavior of employees and its antecedents after an announcement of formal organizational controls.

In terms of security compliance, some scholars focus on the organizational level and others pay more attention to individual level. As to the former, Wall et al. [61] introduced a selective organizational rule violation model into organizational privacy and security contexts and proposed a selective organizational information privacy and security violation model. Sen and Borle [50] examined some public policies, such as public disclosure of vulnerabilities, IT security investment, and data breach laws, which would influence the data breach risk for a state and for organizations within an industry. Parks et al. [47] introduced a theoretical framework that explains the process by which the intended and unintended consequences of implementing privacy safeguards impact organizational privacy compliance. Angst et al. [3] examined whether the way of regulation rule adoption (i.e., symbolic and substantive) had a moderation effect on the relationship between IT security investments and follow-up data security breaches. Lee et al. [35] investigated how firm security would be

influenced by a government's standard, especially with verifiable and unverifiable controls on security.

For the individuals' compliance to ISP, Li et al. [36, 37] identified extrinsic and intrinsic motivation for users' compliance to internet use policy (IUP). Moody (2015) proposed a new integrated model to understand the motivations for employees to accept new ISPs and react negatively against them. Sojer et al. [53] explored drivers of unethical programming behavior in individuals. Chatterjee et al. [10] developed a considerably thorough model to understand unethical IT use from different perspectives of individual, philosophy, sociology, economics, and technology. Boss et al. [9] extensively reviewed protection motivation theory (PMT) and its conventional practice in ISec research to identify opportunities for potential theoretical and methodological improvements on which to build this literature. Lowry et al. [42] explained the behavior of employees to blame organizations and even retaliate against them upon being informed of enhanced ISPs. Hu et al. [23] examined why individuals intentionally violate ISPs via a new paradigm with event-related potentials (ERPs). Posey et al. [48] researched the effect of insiders' organizational commitment levels on threat coping behavior and considered the interconnection of threat and coping appraisal via perceived response cost. Foth [16] explored the factors influencing the intention to comply with data protection in hospitals. Warkentin et al. [65] examined what insiders experience neurologically when faced with fear appeals. Jenkins et al. [25] conducted a behavioral experiment using fMRI and found that alerts in personal computing should be bounded in their presentation, which would cause interruptions to users and make them disregard the alert. Anderson et al. [1] examined the way of habituation to security warning development in the brain through fMRI. Anderson et al. [2] used a type of cognitive neuroscience method called Neuro IS to explore user response to security messages. Johnston et al. [28] identified key factors to explain employees' intention to violate ISPs from the perspectives of disposition and situation.

In terms of security risk management, Wang et al. [63, 64] characterized and distinguished different IS threats in terms of their risk characteristics and further explored the relation of risk characteristics to public searches for information on IS threats. Wright et al. [69] explored why certain influence techniques are especially dangerous when used in phishing attacks. Vance et al. [59] explored an accurate security risk perception measurement and its relationship with security behavior. Oetzel and Spiekermann [46] adopted a privacy impact assessment method to consider privacy issues systematically. Kim and Kim [32] examined how developers of security software learn from managing malware problems. August et al. [4] developed an understanding of how a software vendor approaches the versioning problem and how consumers separate across product variants to diversify security risk when both software as a service (SaaS) and on-premise versions are available. Mitra and Ransbotham [43] explored the relationship between two types of information disclosure (i.e., full and limited) and the diffusion of ISec attacks. Wang et al. [63, 64] showed how application risk from illegal access of insiders could be foreseen through application characteristics. Han et al. [19] investigated the critical antecedents that motivate students to comply immediately with messages from campus emergency

notification systems. Zahedi et al. [71] explained how user's reliance on detection tools is influenced by the performance and cost of the tools. Guo et al. [18] explored the propagation process of malware with a structural risk model. Wolff [68] found that defenses could cause the opposite effect, which exposes the protected systems to new and unpredicted vulnerabilities in the context of complex computer systems. Hui et al. [24] examined whether deterring distributed denial-of-service (DDOS) attacks could be decreased by implementing convention on cybercrime.

4.2 Personnel Security

Personnel security relates to the workforce of an organization. Assessment factors include awareness training, code of conduct, and employment life cycle management. The issue of ethics in security varies from behavioral research to building a network infrastructure. Personnel plays an important role in establishing and maintaining ISec within an organization. Unless IT usage is frequently trained and security awareness and organizational code of conduct are promoted, personnel can inadvertently introduce threats into the organization.

In terms of the external threat targeted on employees, Wright et al. [69] explored why certain influence techniques are especially dangerous when used in phishing attacks to employees. Ho et al. [21] demonstrated the use of different language–action cues of deceivers in different contexts. Wang et al. [63, 64] characterized and distinguished different IS threats in terms of their risk characteristics and further explored how risk characteristics related to public searches for information on IS threats.

In terms of the violation behavior, Sojer et al. [53] explored what drove unethical programming behavior in individuals. Chatterjee et al. [10] developed a considerably thorough model that illustrates unethical IT use from different perspectives of individual, philosophy, sociology, economics, and technology. Hu et al. [23] examined why individuals intentionally violate ISPs through a new paradigm with ERPs. Liang et al. [40] examined and validated several characteristics of malicious insiders noted in the extant literature. D'Arcy et al. explored how complicated and unclear ISec [14] requirements could cause "security-related stress" to employees. Johnston et al. [28] identified key factors to explain employees' intention to violate ISPs, considering disposition and situation. Foth [16] explored the factors that influence employees' intention to comply with data protection in hospitals. Li et al. [36, 37] identified extrinsic and intrinsic motivation for users' compliance to IUP. Lowry and Moody [41] suggested a new integrated model to understand employees' motivations to accept new ISPs and react negatively against them. Lowry et al. [42] explained the behavior of employees to blame organizations and even retaliate against when they are informed about enhanced ISPs. Anderson et al. [1, 2] used Neuro IS, a type of cognitive neuroscience method, to explore user response to the security messages. Anderson et al. [1] used fMRI to examine how habituation to security warnings develops in the brain. Jenkins et al. [25] conducted a behavioral experiment to explain why

individuals would disregard alerts in personal computing. Khansa et al. [30] investigated employees' cyberloafing behavior and its antecedents after an announcement of formal organizational controls.

In terms of the coping mechanisms, Twyman et al. [58] proposed an autonomous scientifically controlled screening system and examined its detection function on individuals' purposely hidden information. Hsu et al. [22] clarified and examined the importance of extra-role behaviors and social controls in employees' compliance to organizational ISP. Johnston et al. [29] explored the effectiveness of an enhanced fear appeal rhetorical framework eliciting a compliance response significantly greater than that produced by contemporary usage of fear appeals. Vance et al. [60] found that the perceived accountability of users could be heightened by the UI design of broad-access systems and further reduced their intentions to violate access policies. Wang et al. [63, 64] showed how to foresee applications' risk from illegal access of insiders through the applications' characteristics. Tsohou et al. [57] proposed a framework to help security managers design and implement ISec awareness programs by treating security awareness as a change process. Posey et al. [48] investigated the effect of organizational commitment levels of insiders on threat coping behavior, considering the interconnection of threat and coping appraisal via perceived response cost. Warkentin et al. [65] examined the neurological experience of insiders when faced with fear appeals. Steinbart et al. [55] focused on what influences users' (dis)continuance of adopting security behavior from the perspective of UI design. Wang et al. [62] examined the coping response mechanism of employees in the process of phishing email detection.

4.3 Threat Mitigation

Threat mitigation is concerned with network segmentation (e.g., network security infrastructure, intrusion detection, and remote access), vulnerability management (e.g., scanning, patching, and standard operating procedures), content checking (e.g., data filtering and virus protection), and incident management issues (e.g., forensics and event correlation). Related reviewed papers are summarized as follows.

In terms of vulnerability management, Wright et al. [69] explored why certain influence techniques are especially dangerous when used in phishing attacks, which helped identify this type of vulnerability. Guo et al. [18] explored the propagation process of malware using a structural risk model. Sen and Borle [50] examined some public policies, such as public disclosure of vulnerabilities, IT security investment, and data breach laws, would influence the data breach risk for a state and for organizations within an industry. Chen and Zahedi [11] considered the effect of cultural difference on security behaviors based on context-sensitive theory. Li et al. [36, 37] found that organizational justice and personal ethics are two effective levers to mitigate the risk of violation of IUP. Wang et al. [62] suggested that emotion-focused coping of employees and avoidance coping in phishing email coping will cause vul-

nerability. Jenkins et al. [25] found that alerts pervasive in personal computing will create vulnerability if they are not bounded in their presentation.

In terms of content checking, Zahedi et al. [71] explained how user's reliance on detection tools is influenced by their performance and cost. Ho et al. [21] demonstrated that deceivers would use different language–action cues in different contexts. Siering et al. [51] derived different linguistic and content-based cues that were used as input for various fraud detection classifiers. Li et al. [38] developed advanced text mining techniques to analyze multilingual textual traces in underground economy and identify key international underground economy sellers. Liang et al. [40] examined and validated several characteristics, which could be used to identify malicious insiders.

In terms of incident management, Goode et al. [17] explored how a breached organization could best determine the optimal level of customer compensation in response to data breach, which is about incident management. Choi et al. [12] developed a model to show how firms' recovery measures influence customers' behavior online after data breach. Angst et al. [3] examined whether the manner in which regulation rules were adopted (i.e., symbolic and substantive) had a moderation effect on the relationship between IT security investments and follow-up data security breaches. Jensen et al. [26] found that participants who received mindfulness training could better avoid the phishing attack that those who did not. Mitra and Ransbotham [43] explored the relationship between two types of information disclosure (i.e., full and limited) and the diffusion of ISec attacks.

4.4 ISec Economics

Some scholars have argued that the focus of IT security management is shifting from what is technically possible to what is economically efficient. ISec economics refers to using economic theory in handling ISec decisions, such as the ISec investment and consumer choice. The former is about how an organization makes decision on ISec investment based on the return and loss without investment. The latter explores how transaction security can be enhanced from the economics perspective to increase the transaction intention of consumers. Related reviewed papers are summarized as follows.

Lee et al. [35] investigated how firm security would be influenced by a government's standard, especially when verifiable and unverifiable controls on security concerns are available. August et al. [4] explained how a software vendor approaches the versioning problem and how consumers separate across product variants to diversify security risk when both SaaS and on-premises versions are available. Choudhary and Zhang [13] explored the impact of a change in the distribution of defect-related costs on the release time of vendors and patching strategy under SaaS. Sen and Borle [50] examined some public policies (e.g., public disclosure of vulnerabilities, IT security investment, and data breach laws) that would influence the data breach risk for a state and for organizations within an industry. Ji et al. [27] analyzed a size-based

security monitoring policy with and without profiling. Jensen et al. [26] developed a novel security training method given that some employees are used to training based on rules. Goode et al. [17] studied how a breached organization could best determine the optimal level of customer compensation in response to data breach.

4.5 Privacy

In this study, information privacy is assigned to the area of using privileged information with malicious intent, which includes the following parts: (1) Policy, practices, and controls. This part includes development of taxonomies, as well as rule definitions, impact assessments, and awareness and training: (2) Privacy and information management strategy. This assessment includes description of privacy information strategies, requirements, and compliance processes, as well as incident response situations: (3) Data, rules, and objects. This part includes the development of classification and/or business process models. Related reviewed papers are summarized as follows.

Li and Sarkar [39] proposed a dynamic value-concatenation method for data privacy protection and data quality preservation for application. Oetzel and Spiekermann [46] adopted a privacy impact assessment method in considering privacy issues systematically. Wall et al. [61] introduced a selective organizational rule violation model into the contexts of organizational privacy and security. Parks et al. [47] evaluated the intended and unintended consequences of implementing privacy safeguards and their impacts on organizational privacy compliance. Goode et al. [17] investigated how a breached organization could best determine the optimal level of customer compensation in response to data breach. Choi et al. [12] developed a model showing how firms' recovery measures influence customers' behavior online after data breach. Angst et al. [3] examined whether symbolic and substantive adoption would moderate the effect that IT security investments had on reducing the incidence of data security breaches over time.

4.6 Transaction and Data Integrity

Transaction and data integrity is concerned with business process transaction security (e.g., fraud detection and transaction security), database security (e.g., configuration and control), message protection (e.g., encryption and message security), secure storage (e.g., data storage, archiving, retrieval, and destruction), and system integrity (e.g., secure system management and business continuity planning). Six papers we reviewed have been identified as the following.

Li and Sarkar [39] proposed a dynamic value-concatenation method that could protect data privacy while preserving data quality for application. Kim et al. [31] indicated that cultural difference is an important element in designing e-commerce

websites and security protection for multinational companies for a worldwide audience. Bhattacherjee and Park [7] explained why users move from client-centric computing to cloud computing. Herath et al. [20] explored users' intention to adopt an email authentication service. Siering et al. [51] derived different linguistic and content-based cues used as input for various fraud detection classifiers, which helped identify the fraud. Li et al. [38] developed advanced text mining techniques to identify the key sellers in Cyber Carding Community.

4.7 Identity and Access Management

This part includes identity proofing through background screening and alternative methods of credential management. Identity access management focuses on identifying users, protecting confidential information from unauthorized users, and providing authorized users secure and controlled access to resources. Related reviewed papers are summarized as follows.

Steinbart et al. [55] showed that poor performance (login failures) of identity authentication resulted in discontinuance of a secure behavior and the adoption of less-secure behaviors. Vance et al. [60] found that the perceived accountability of users could be increased by the UI design of broad-access systems and further reduced their intentions to violate access policies. Roßnagel et al. [40] examined whether individuals would like to pay when using federated identity management. Herath et al. [20] investigated users' intention to adopt an email authentication service.

4.8 Application Security

An application security assessment entails code review, secure coding practices, and secure policies and procedures to manage SDLC (Systems Development Life Cycle). Preventing a security error is generally less costly than fixing it once it occurs. SDLC includes procedures that ensure security throughout its process. In papers we reviewed, only one pertained to application security. Sojer et al. [53] explored the drivers of unethical programming behavior by individuals in the processes of systems development.

4.9 Physical Security

Security is not an issue that can exclusively be handled with software. Requirement for physical barriers exists as well. Physical security describes the measures taken to protect facilities from potential attackers. In the IBM model, two topics are considered, namely, site management and physical asset management. No reviewed

papers have discussed this theme, which can be explained that this topic related to other disciplines. Furthermore, other physical security concerns will emerge with the development of artificial intelligence. IoTs will force various intelligent devices, which may expose organizations to new vulnerabilities. For this reason, we retained this theme in our discussion.

5 Recommendations

To some extent, the identification of research streams based on IBM security model can help us learn about the relationship of ISec research and industry requirements. However, the next steps for industry and research communities still need further study. Thus, on the basis of existing literature, we focused on four objects in organizations, namely, data, human behavior, business processes, and IT/IS, as shown in Fig. 3. From the dynamic view, we considered the interaction between two objects. An organization can make policies to manage each flow to minimize security risk. Based on this framework, some main findings in existing research are summarized, and the recommendations for industry and research communities are discussed and provided.

We will explain each object and then consider the flow or interaction between them. First, we define data as the core object in the process of business digitization and therefrom ISec management. Data of organizations include the internal and external data that mainly originate from two sources, namely, human behavior related and business processes related, such HER of patients [33], online customer behavior data [12], and organizational operation and financial data [35]. Organizations should consider using which type of IS/IT to record, store, transfer, and protect data while

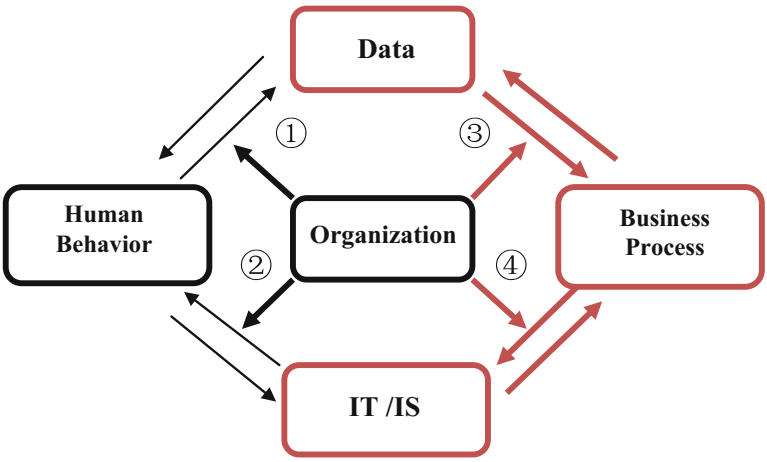


Fig. 3 Proposed ISec research framework in the process of business digitization

restraining human behavior (insiders and consumers) and encrypting the business process to protect data security; thus, these three objects are considered as another three objects.

Determining the IT/IS that supports and accomplishes the business process (e.g., OA, ERP, and CRM) with the IT/IS protection security (e.g., firewall technology, intrusion detection system, proxy servers, and virtual private networking) is necessary. The former is the infrastructure of digital businesses, and its vulnerability raises new threats to organizations [43]. Hence, the latter is designed and deployed to protect the former. Data are flowing in the system, and insiders should follow standard system usage.

Two main types of human behavior are considered by organizations, namely, insiders and consumers (users). The former is considered among the greatest threats to organizational ISec [42]. Insiders' violation to ISPs has caused great loss to organizations, and consumers' insecurity behavior also expose organizations to the threats. Moreover, consumer behavior will generate the data collected and analyzed by organizations to make managerial decisions, which is the main component of digital businesses. However, the illegal usage of collected behavioral data is an increasing concern for consumers.

With regard to the business processes considered as the core object in this study, we adopted the definition from Wikipedia, "Business process is a collection of related, structured activities or tasks that produce a specific service or product (serve a particular goal) for a particular customer or customers." This definition is closer to that of operational processes rather than the management and supporting processes. Operational processes are the core business and realize the primary value stream, such as opening an account in a bank after taking orders from customers. However, management processes, which mainly include "corporate governance" and "strategic management" govern systems operation. Supporting processes include health and safety, accounting, recruitment, call center, and technical support.

The development of organizational business process management (BPM) involves three stages. The first stage started with advances in the data-driven methods. In this stage, data storage and retrieval technologies have made great progress. Most of the IS was developed with data-modeling method; thus, BPM had to adapt to the system neglecting business processes. BPM reached the second stage with the emergence of various enterprise resource planning software in the 1990s, when the business-process-oriented management system increasingly dominated the market. In the third stage, e-businesses, which benefitted from the development of internet technologies, realized the automated business processes across organizations. This automated process created a platform that integrated sellers and buyers online and promoted collaboration and integration of people, systems, processes, and information within and across enterprises.

Currently, some newly emerged information technologies, such as cloud computing, social network, mobile technology, and big data analytics, are reforming BPM. For example, cloud computing technology has significantly increased the computing power of companies with low cost and has eliminated the restraint of location. Social media and smartphones have produced new channels for organizations to reach their

Table 4 Summary of recommendations for information security research

Directions	Suggestions for researchers	Suggestions for organizations	Related themes in IBM security model
Interaction between data and human behavior	1. Before data breach, how to encourage compliance from the perspective of positive employee 2. After data breach, how to conduct repair strategy	1. Watching out for the side effect of enhanced ISPs 2. Balance between compensation and over-compensation	Governance; Personnel security; Privacy
Interaction between human behavior and IT/IS	1. For employees: emotions' effect on unethical IT usage behavior 2. For consumers: dual effect of IT/IS used for security protection on transaction behavior	1. Pay more attention to users' emotional management 2. Balance between security protection measures and interference	Governance; Personnel security; Threat mitigation; Identity and Access management
Interaction between IT/IS and business process	1. Data analytics technologies used for business fraud detection 2. Information security economics used for security investment researches	1. More IT security investment doesn't mean more security, especially in complex system	Governance; Application security; Physical security; Information security economics
Interaction between business process and data	1. How to jump out of the dilemma: privacy protection and personalized service 2. How to choose a suitable data view: right or commodity	1. New data analytics method need to be developed 2. Cross-order transaction should consider the different privacy view hold by two trading parties	Governance; Transaction and data integrity

customers. Various customer data collected through these channels have led to a significant growth in business analytics based on big data technologies, which can help managers to make effective managerial decisions and serve their customers well.

To sum up, business processes must reconsider its relationship with IT/IS and data. In the following, we proposed four directions by examining the interactions between different objects to identify the potential research opportunity for subsequent researchers and provide some suggestions to industry community based on 59 reviewed papers in this chapter (Table 4).

5.1 Interaction Between Data and Human Behavior

For the interaction between data and human behavior, data generation based on consumer behavior and data protection conducted by employees are two important themes to consider. For data generation, consumers worry about data security and they may give up accepting services especially when they do not believe the organization. Kohli and Tan [33] claimed that patients' privacy calculus may impact their EHR data sharing. Failure to protect consumer data and prevent privacy breaches can cause great damage to the reputation and finances of a company [46]. Considering the leaked behavior of organizations, except for organizational malicious leaked behavior, Wall et al. [61] demonstrated that organizations intentionally choose to violate users' data security protection rules required by the government to decrease the excessive cost of safety protection. With the increasing cases of privacy breach, some researchers have begun to explore how a breached organization can decide on customer compensation after a data breach and retain their consumers [12, 17].

For data protection, employees' compliance to ISPs has been an extremely popular topic. Among the 59 papers we reviewed, 28 papers have referred to this topic [9, 14, 16, 22, 28–30, 36, 37, 40, 41, 48, 57, 62, 65]. Data breach resulting from insiders' security policy violation behavior can be seen as the unintentional leaked behavior of organizations. Although organizations often develop various policies to restrain employees' behavior [16], most of the security losses are caused by such behavior.

5.1.1 Suggestions for Researchers and Organizations

Before data breach: Paying attention to the side effect of enhanced ISPs
ISPs can significantly improve organizational security situation. However, researchers are increasingly realizing that excessive ISPs has brought "security-related stress" to employees [14]; thus, employees react negatively and even retaliate against the organization by intentionally violating enhanced ISPs [41, 42]. Some scholars have explored solutions to this phenomenon. Balozian and Leidner [6] argued that well-justified security additions are useful to improve employees' attitude to enhanced ISPs. Hsu et al. [22] suggested that the extra-role behavior of employees should be emphasized and that social control from colleagues can complement the formal ISPs of organizations. Researchers have set an inappropriate assumption about negative employees. Organizations keep increasing their ISPs because they think that employees often attempt to violate policies. On the basis of the effect of enhanced ISPs, researchers can eliminate this assumption and explore how to encourage employees' compliance to ISPs from the perspective of positive employees. Organizations should be careful in introducing new ISPs by using suitable methods. Employees will be irritated and react negatively if managers ignore their rights and freedoms as humans when introducing potentially freedom-restricting policies [41]. Respect and fairness are two basic factors that organizations should show to employ-

ees [42]. If necessary, providing sufficient justification of the enhanced policies is suggested.

After data breach: more attention to repair strategies
When data breach happens, organizations have to respond to it even if a great damage has already occurred. Existing service failure literature has shown that effective repair strategies play an important role in retaining consumers and improving their repurchase intention. Therefore, exploring effective repair strategies after data breach is necessary. In our reviewed papers, only two papers have focused on this question [12, 17]. Based on Sony's data breach case, Goode et al. [17] presented an adapted model to explain customer responses to a data breach recovery action. Organizations should provide compensation depending on breach severity; however, overcompensation is not a good idea. With the advent of social media, researchers have been interested in exploring how customers spread comments (positive and negative) in their social network after data security breaches and how spreading of data breach will influence organizational compensation packages and outcomes.

5.2 Interaction Between Human Behavior and IT/IS

With regard to the interaction between human behavior and IT/IS, we initially considered the security concerns in the process of IT/IS usage. In this perspective, we identified the unethical IT usage behavior in the literature. Four reviewed papers discussed this topic [10, 30, 53, 55]. However, IT/IS can be also regarded as tools to be used to detect malicious user behavior and monitor employees' behavior. In this perspective, developing a detection system has been a popular research topic [21, 23, 58], including UI design [1, 2, 55, 60]. In addition, Neuro IS technologies are increasingly adopted by IS scholars [1, 2, 25].

How to adopt countermeasures according to different unethical IT/IS usage behaviors?
To decrease the vulnerability of organizational IS, taking effective measures is necessary to respond to different unethical behaviors of employees. Unethical IT usage behavior is complex, and Chatterjee et al. [10] found that it would be influenced by a wide range of individual, philosophical, social, economic, and technological factors. Three different unethical IT usage behaviors, namely, malicious, intentional but not malicious, and unintentional, are identified to address unethical IT usage behavior effectively. For unintentional unethical IT usage behavior, enhancing user's risk awareness of their behavior by SETA programs is necessary [57]. For example, cyberloafing behavior of employees is a representative type of behavior; it exposes the organization to internet threats. Khansa et al. [30] found that cyberloafing behavior is mainly influenced by past tendencies to cyberloaf and others' influence. Organizations can significantly decrease this behavior by adopting formal controls (e.g., penalty). Steinbart et al. [55] also found that employees' habits on technology usage in daily life would carry over to the workplace, which shows the importance of secu-

rity awareness training. For intentional but not malicious unethical IT usage behavior, sanctions are more effective [10]. For example, for unethical programming behavior as a representative of this type of behavior, Sojer et al. [53] explored its drivers and encouraged firms to prevent it by informing developers of its negative consequences. Finally, for malicious unethical IT usage behavior, identifying and minimizing it by security training or sanction controls is difficult, especially when users pretend to follow organizational ISPs. To address these problems, some scholars have attempted to develop or adopt different IT/IS to detect this behavior. This topic is discussed as follows.

How to develop/adopt IT/IS to deal with security concerns related to human behavior?

The perspective of IT/IS as infrastructure supporting business process has changed to IT/IS as tools identifying insecurity behavior or enhancing security behavior intention. For example, Twyman et al. [58] proposed autonomous scientifically controlled screening systems that can detect information hidden by individuals. Hu et al. [23] proposed a new paradigm based on event-related potentials that can be used to identify individual violations of ISPs based on their self-control difference. Moreover, Ho et al. [21] found that specific language–action cues influenced by context can be used to identify computer-mediated deception. In addition, Neuro IS technologies, such as fMRI and eye movement-based memory, have been adopted by scholars to learn individuals' real intention based on their physiological change when exposed to security risk or warning/alerts of organizations [1, 2, 25, 59]. Aside from using IT/IS to identify insecurity behavior, UI design is developed to help users continue to adopt security behavior [55] and not to violate access policies [60]. This complementary measure is suggested because it can target repetition suppression in users' brain, such as the polymorphic warning, which can elicit positive effect in milliseconds without additional cost [1].

5.2.1 Suggestions for Researchers and Organizations

For employees: effect of emotions on unethical IT usage behavior

Most of the existing studies are from the perspective of rational behavior, which is based on PMT [9, 48] and deterrence theory [16, 25] without fully considering the emotion of individuals, such as rage, anger, and despair. However, an increasing number of scholars have realized that the emotional state of individuals is an important foundation for them to make rational decisions. Formal sanctions will be effective for individuals who perceive low to moderate level of anger, but neither formal nor informal sanctions will lose efficacy on individuals who perceive high level of anger. Willison and Warkentin [67] indicated that a new stream of research for the IS security field is to examine the relationship between emotions and deterrence. For example, will organizational injustice result in negative emotions or will this emotion further influence individuals' unethical IT usage behavior as revenge [42]? Do emotions moderate the effect of threat of sanctions on unethical IT usage

behavior or does the extent of emotions play different roles? Organizations should pay more attention to employees' emotion management to ensure the efficacy of the ISPs, especially with sanctions.

For consumers: dual effect of IT/IS used for security protection on transaction behavior

Although enhanced security protection for consumers is adopted by organizations, unexpected results are identified. Kim et al. [31] found that perceived effectiveness of web assurance seal services (WASS) from organizations would influence the transaction intention of American consumers. Steinbart et al. [55] claimed that UI design in mobile paradigm would influence login success rates, which would further result in consumers' discontinuance of a secure behavior. When consumers feel that their consumption is interrupted by over-security measures, they may become impatient and discontinue shopping. Jenkins et al. [25] conducted behavioral experiment using fMRI and found that the presentation of alerts should be carefully controlled because the timing of interruptions strongly influences alert disregard. To sum up, existing research actually requires a balance between IT/IS used for security protection and interference brought by these IT/IS. Although security protection is beneficial to consumers, consumers may still perceive disturbance by excessive and fussy authentication. This phenomenon can be explained by the "dual-task interference" (Anderson et al. [1], which indicates that multitasking is difficult for people. Consumers find it difficult to shop while passing security validation. Thus, designing some IT/IS to protect consumers' security with least interference is challenging. In addition, new IT may bring more complexity to manage consumers' security behavior. For example, since the emergence of mobile technology, more consumers use mobile devices to shop instead of their computers. However, security policies effective on desktop computing paradigm will not work in the mobile paradigm [2].

5.3 Interaction Between IT/IS and Business Process

New emerging IT/IS will be adopted to support business processes at the cost of new channels of vulnerabilities to be exposed to security threat, such as cloud computing [7]. However, new business processes or models based on new IT/IS also trigger new types of attacks, such as business fraud [51], phishing [62, 69], malware propagation [18], and underground economy sellers [38], which urge matched IT/IS investment to provide security protection.

5.3.1 For New IT/IS-Enabled Business Process

Security concerns emerge mainly depending on the characteristics of new business processes or models. Based on reviewed papers, we show different security concerns that occur in different IT/IS-enabled business processes. The first is about the cloud

computing service. Cloud service can provide users with universal access to cloud-hosted resources and processing power with low cost [7]. However, the cloud service provider will easily attract and receive denser attacks from hackers. For example, SaaS is a type of business application of cloud computing. August et al. [4] pointed out that the SaaS versioning of software has relatively higher directed risk than the traditional on-premises version because one vulnerability of the SaaS is letting a malicious attacker affect many organizations using this SaaS all at once. Second, for crowdfunding platforms that provide possibility for project realization even with lack of fund, their drawback is the rising risk of fraud related to the project campaigns prevalent on these open online services. Given that project founders often only have project ideas without the actual product during the funding period, judging the legitimacy of the project is difficult [51]. Third, for underground economy, such as Cyber Carding Community, the development of internet technologies and illegal business application of internet also call for solutions. In addition, cross-border e-commerce websites can realize the transaction among different countries online, which greatly reshapes the international business model. However, the cultural differences should be considered to design website authentication in this context. Kim et al. [31] found that the effectiveness of WASS influences transaction intention of US consumers but not Korean consumers. Email has become a daily used business communication software. However, email phishing attack has caused great loss to organizations [62]. Herath et al. [20] explored how to increase the adoption of an email authentication service by controlling this risk to organizations. Finally, malware propagation is also one top security challenge in business processes [18].

5.3.2 For Security Investment on IT/IS for New Business Process

New business processes emerging with new IT/IS expose organizations to new security risk. Therefore, organizations must adopt enhanced or targeted security protection measures. One important topic in this part is organizational security investment decision on IT/IS. When considering security investment, rules of the government will have some restraints on organizational decisions. Angst et al. [3] found that the effectiveness of IT security investments would be weakened by symbolic adoption of government rules and further increase the risk of data breach in business processes. The notion of buying more (and even more expensive) defense technologies and systems is held by many organizations. Companies think that the quantity of security protection technologies will increasingly improve their security of business processes. However, Wolff (68) claimed that more is not always better, especially in defending a complex system. New and unpredictable vulnerabilities will be produced by interactions among different components of system and defense mechanisms. Adding defenses to this type of complex system can actually undermine its security.

5.3.3 Suggestions for Researchers and Organizations

Data analytics technologies used for business fraud detection
IT/IS-enabled business processes show all types of new security threats faced by organizations, especially business fraud detection. To fill this gap, data analytics technologies should be adopted. In our reviewed papers, some scholars have attempted to pioneer. For example, data mining method was adopted by Siering et al. [51] to detect the fraudulent behavior on a crowdfunding platform. Results showed that different linguistic and content-based cues can be used to identify fraud in business processes. Li et al. [38] also developed a novel system using advanced text mining techniques to analyze multilingual textual traces in the underground economy and further identify key sellers. Guo et al. [18] conducted an analysis on the propagation process of malware with social network data. As suggested by Li et al. [38], the question of how to use hacker community data to inform cybersecurity intelligence remains open as hackers increasingly congregate in their communities. Leveraging social media analytics to probe into business fraud awaits further exploration.
ISec economics used for security investment research
Based on the above discussion on security investment of organizations, one important potential research direction is the perverse effects of security investment. Although, taxonomy for the sources of different perverse effects in security has been proposed by Wolff [68], several questions have been opened. What types of defenses cause these effects in practice and why? What is best action to avoid or counteract them? Future research in this area can further elaborate Woff's understanding of when and why perverse effects arise in defending computer systems in business processes and how they may be most effectively mitigated. Organizations' new security investment to protect IT/IS should be reviewed not only for their individual impact but also for their interactions with other system components and usability features.

5.3.4 Interaction Between Business Process and Data

The final and most important type of interaction we considered is the interaction between business process and data. This interaction produces at least two themes. One is the data-driven business process, and the other is data generation and protection in business processes.

5.3.5 For Data-Driven Business Processes: A Dilemma

With data increasingly used for managerial decision, traditional business processes, such as product design, marketing, and customer management, have been reshaped by data and reached the digital business. However, data-driven business processes also bring new challenges to organizations. One important topic is how to keep the balance between privacy protection and personalized service based on analysis of personal data. For example, recommended systems are adopted by an increasing number of

companies to analyze the demand of customers and then recommend goods or service to customers. However, whether consumers will feel invaded when they received the recommendation has been one difficult question to answer. Privacy-related paradox has been noticed by some scholars [5, 34, 54, 56], that is, organizations face a dilemma where consumers want to share the benefit of data-driven recommendation but not willing to share their data because of privacy concerns.

5.3.6 For Data Protection in Business Process: Illegal Data Usage and Data Breach

Based on the dilemma discussed above, data-driven business processes have the potential orientation to privacy invasion. For organizations, all types of data protection technologies are required to be deployed to increase consumers' trust and respond to government rules [31]. The effectiveness of IT investment has been discussed in the part of "interaction between IT/IS and business processes." Similarly, if data breach is detected, organizations also need to provide compensation to consumers, as mentioned in the first type of interaction. In this part, we discuss the effect of business decision on data sharing with the restraint of government rules. For example, Mitra and Ransbotham [43] focused on organizational decision on information disclosure of vulnerability and found that full disclosure would lead to greater risk than limited disclosure. Furthermore, Sen and Borle [50] found that the risk of data breach would be significantly influenced by the strictness of laws on data breach disclosure. Moreover, when adopting the commercial perspective of the concept of privacy, privacy will be seen as a type of goods to be traded [54]. For example, when considering the data transforming healthcare, privacy calculus is argued to influence patients' data sharing [33].

5.3.7 Suggestions for researchers and organizations

How to jump out of the dilemma: Innovation on data analytics method
The above discussion on data-driven business processes show the dilemma organizations face. Potential solution still needs to be determined from data analytics itself. For example, regression tree is a type of data analytics method but it can also be used as a tool for mining personal information, such as regression attacks [39]. To address this problem, Li and Sarkar [39] developed a new dynamic value-concatenation method approach. This approach can ensure the quality of data while avoiding privacy infringement. Therefore, this type of data analytics methods should be further studied although it may take a long for researchers. This type of method will encourage customers to share their data without worrying about privacy and then share the benefit from the personalized service.

How to choose a suitable data view: Effect of cultural difference on privacy concerns in business process

Privacy calculus is another way to deal with the organizational dilemma because organizations can buy the personal data when consumers accept the notion of privacy as goods. Thus, an increasing number of scholars has been interested in privacy calculus to some extent. However, one underlying question is whether considering cultural difference is reasonable and applicable. This topic is meaningful with the development of cross-border e-commerce and cross-national companies, especially in the management discipline. Smith et al. [54] summarized two value-based privacy views, namely, privacy as a right and privacy as a commodity, and presented the dissonance between US and European privacy laws. Europe tends to see privacy as more of a property right by consumers compared with the US. When cross-border transaction occurs between countries in Europe and the US, problems may emerge without fully considering the cultural difference. Therefore, further research, which focuses on this question and explores guidelines for organizations, is suggested.

6 Limitations

First, although we have explained that we made a conscious decision to prioritize quality over quantity, we only considered six journals. Thus, more journals are encouraged to be considered. Second, although we analyzed each paper carefully according to the assessment of IBM themes, some subjective assignments are admitted to be inevitable. Moreover, the security model of IBM selected as the representative of industry requirement may ignore some ISec themes or consider other themes (e.g., physical security) that are closer to other disciplines, such as computer science. Hence, determining a suitable security model of industry is encouraged. In addition, we only considered four types of interaction of the four objects because of their relative importance. Other types of interaction also deserve consideration, especially the interaction among three of four objects. Finally, given that big data analytics have developed mainly in the past five years, we only reviewed papers published from 2014. However, expanding the term of our review is also encouraged.

7 Summary and Conclusions

In this chapter, we reviewed ISec research published in MISQ, ISR, JMIS, JAIS, EJIS, and ISJ from 2014 to 2107 and coded each paper into one or more themes of IBM security model. Then, we evaluated the relationship between ISec academic research and ISec industry requirement. Some increasingly popular themes, such as privacy, threat mitigation, and transaction and data integrity, for IS researchers were specified, and four objects related to ISec in organizations were identified. By further coding each paper into one or more objects, we considered the interaction between two objects. Based on each type of interaction, some suggestions for IS researchers and organizations were provided. Based on the topic of this book chapter, we strongly

recommend that researchers and organizations pay more attention to the interaction between IT/IS and business processes, and interaction between business processes and data. Both these interactions represent the process of business digitization, from which some security topics are worth to be further explored, especially in the digital era.

Appendix

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper

Articles	Themes of IBM security model								
	Governance	Privacy	Threat mitigation	Transaction and data integrity	Identity and access management	Application security	Physical security	Personnel security	Information security economics
Total (59)	50	7	18	6	4	1	0	28	8
1. Wright et al. [69]	√		√					√	
2. Lee et al. [35]	√								√
3. Hsu et al. [22]	√							√	
4. Wang et al. [63, 64]	√							√	
5. August et al. [4]	√								√
6. Steinbart et al. [55]	√				√			√	
7. Mitra and Ransbotham [43]	√		√						
8. Choudhary and Zhang [13]	√								√
9. Johnston et al. [29]	√							√	
10. Han et al. [19]	√								
11. Kim and Kim [32]	√								
12. Li and Sarkar [39]		√		√					
13. Vance et al. [60]	√				√			√	

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper									
Articles	Themes of IBM security model								
	Governance	Privacy	Threat mitigation	Transaction and data integrity	Identity and access management	Application security	Physical security	Personnel security	Information security economics
Total (59)	50	7	18	6	4	1	0	28	8
14. Chen and Zahedi [11]	√		√						
15. Wang et al. [63, 64]	√							√	
16. Boss et al. [9]	√							√	
17. Guo et al. [18]	√		√						
18. Ho et al. [21]			√					√	
19. Wolff [68]	√								
20. Hu et al. [23]	√							√	
21. Chatterjee et al. [10]	√							√	
22. Posey et al. [48]	√							√	
23. Sen and Borle [50]	√		√						√
24. Twyman et al. [58]								√	
25. Sojer et al. [53]	√					√		√	
26. D'Arcy et al. [14]	√							√	
27. Johnston et al. [28]	√							√	
28. Oetzel and Spiekermann [46]	√	√							
29. Tsohou et al. [57]	√							√	
30. Roßnagel et al. [49]					√				
31. Anderson et al. 2016	√							√	
32. Foth [16]	√							√	
33. Kim et al. [31]				√					

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper									
Articles	Themes of IBM security model								
	Governance	Privacy	Threat mitigation	Transaction and data integrity	Identity and access management	Application security	Physical security	Personnel security	Information security economics
Total (59)	50	7	18	6	4	1	0	28	8
34. Siponen and Vance [52]	√								
35. Bhattacherjee and Park [7]	√			√					
36. Warkentin et al. [65]	√							√	
37. Wall et al. [61]	√	√							
38. Zahedi et al. [71]			√						
39. Vance et al. [59]	√								
40. Herath et al. [20]				√	√				
41. Li et al. [36, 37]	√		√					√	
42. Lowry et al. [42]	√							√	
43. Lowry et al. [42]	√							√	
44. Wang et al. [62]	√		√					√	
45. Siering et al. [51]			√	√					
46. Parks et al. [47]	√	√							
47. Niemimaa and Niemimaa [45]	√								
48. Liang et al. [40]			√					√	
49. Li et al. [38]			√	√					
50. Kohli and Tan [33]			√						
51. Khansa et al. [30]	√							√	
52. Ji et al. [27]	√								√

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper									
Articles	Themes of IBM security model								
	Governance	Privacy	Threat mitigation	Transaction and data integrity	Identity and access manage-ment	Application security	Physical security	Personnel security	Information security economics
Total (59)	50	7	18	6	4	1	0	28	8
53. Jensen et al. [26]	√		√						√
54. Jenkins et al. [25]	√		√					√	
55. Hui et al. [24]	√								√
56. Goode et al. [17]	√	√	√						√
57. Choi et al. [12]	√	√	√						
58. Angst et al. [3]	√	√							
59. Anderson et al. [1, 2]	√		√					√	

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper					
Theoretical basis	Methodology method	Objects related			
		Data	Business process	Technology/system	Human behavior
		10	16	24	43
Persuasion theory; motivation theory	Experiment				✓
Game theory	Modeling	✓			
Social control theory	Survey				✓
Information foraging theory	Survey				✓
Microeconomic theory	Modeling		✓		✓
Protection motivation theory (PMT) Technology threat avoidance theory (TTAT)	Experiment			✓	✓
Diffusion of innovation	Modeling		✓	✓	
Microeconomic theory	Modeling		✓		✓
Protection motivation theory, deterrence theory	Interview; Experiment				✓
(Etzioni's) Compliance theory	scenario-based survey				✓
Theory of knowledge retention	Modeling		✓	✓	

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper					
Theoretical basis	Methodology method	Objects related			
		Data	Business process	Technology/system	Human behavior
		10	16	24	43
K-anonymity framework; regression tree	Algorithm; Experiment	✓		✓	
Accountability theory	Factorial Survey			✓	✓

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper					
Theoretical basis	Methodology method	Objects related			
		Data	Business process	Technology/system	Human behavior
		10	16	24	43
Context sensitive theory; TTAT; PMT; coping theory	Survey				✓
Routine activity theory	Modeling				✓
Protection motivation theory (PMT)	Reviewing; Experiment				✓
Social network analysis	Modeling			✓	
Interpersonal deception theory; social distance theory; media richness theory	Algorithm; Modeling			✓	✓
Theory of unintended consequences	Case study			✓	
Self-control theory	Experiment				✓
Theory of planned behavior (TPB); Universal philosophical theories of ethics	Scenario-based survey		✓		✓
Protection motivation theory	Survey				✓
Opportunity theory of crime, Institutional anomie theory; institutional theory	Modeling	✓	✓		
Orienting theory, defensive response theory	IS development			✓	✓
Theory of planned behavior model; expected utility, deterrence, and ethical work climate theory	Survey			✓	✓
Coping theory; moral disengagement theory Social cognitive theory	Survey				✓

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper					
Theoretical basis	Methodology method	Objects related			
		Data	Business process	Technology/system	Human behavior
		10	16	24	43
PMT; general deterrence theory	Survey				✓
Privacy impact assessment (PIA)	Approach Development	✓		✓	
Actor-network theory (ANT), Structuration theory; Contextualism	Action research				✓
Economic theory	Modeling; Experiment			✓	✓
	Experiment			✓	✓
TPB; general deterrence theory	Survey				✓
Cultural dimensions	Survey		✓	✓	✓
	Case study				✓
Migration theory	Survey		✓		✓
Fear appeal theory	Experiment			✓	✓
Selective organizational rule violations model; Strain Theory; general deterrence	Theory development	✓	✓		
PMT; detection tool impact (DTI) theory	Experiment			✓	✓
TPB; context-updating theory Dual-task interference theory	Experiment			✓	✓
TAM; TTAT; PMT	Survey			✓	✓
Justice theory; sanction theory	Survey				✓
Organizational control theory reactance theory	Survey				✓
Fairness theory; reactance theory deterrence theory	Survey				✓
Extended parallel process model behavioral decision-making	Survey			✓	✓

(continued)

(continued)

Appendix: Coded Articles: literature Classification/Theory/Method/Research objects related in each paper					
Theoretical basis	Methodology method	Objects related			
		Data	Business process	Technology/system	Human behavior
		10	16	24	43
Four-factor theory; Leakage theory Verbal cues; Static Dynamic Linguistic Competence model of fraud detection Information manipulation theory Criteria-based content analysis Scientific content analysis Reality monitoring Channel expansion theory Interpersonal deception theory Interaction adaptation theory	Data mining/machine learning		✓	✓	
Grounded theory	Interpretive grounded theory	✓	✓		
Practice theory	Ethnography		✓		
Trait theory	Text mining				✓
Text mining technologies	Experiment; Case study		✓	✓	
	Integration -analytics technologies	✓			✓
Social learning theory	Survey				✓
Traditional monitoring methods	Modeling			✓	
Mindfulness theory	Design science				✓
Functional magnetic resonance imaging (fMRI)	Behavioral experiment			✓	✓
General deterrence theory; routine activity theory	Modeling				
Expectation confirmation theory;	Longitudinal field study of Sony customers	✓	✓		✓
Theory of justice; Psychological contract theory	Scenario-based survey	✓	✓		✓
Institutional theory	A growth mixture model approach	✓	✓		
Habituation theory	Functional magnetic resonance imaging			✓	✓

References

1. Anderson, B. B., Vance, A., Kirwan, C. B., Eargle, D., & Jenkins, J. L. (2016). How users perceive and respond to security messages: A NeuroIS research agenda and empirical study. *European Journal of Information Systems, 25*(4), 364–390.
2. Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., & Eargle, D. (2016). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems, 33*(3), 713–743.
3. Angst, C. M., Block, E. S., D'Arcy, J., & Kelley, K. (2017). When do IT security investments matter? Accounting for the influence of institutional factors in the context of healthcare data breaches. *MIS Quarterly, 41*(3), 893–916.
4. August, T., Niculescu, M. F., & Shin, H. (2014). Cloud implications on software network structure and security risks. *Information Systems Research, 25*(3), 489–510.
5. Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly, 30*(1), 13–28.
6. Balozian, P. Y., & Leidner, D. (2016. December). *IS security Menace: When Security Creates Insecurity*. Paper presented at International Conference on Information Systems, Dublin, Ireland.
7. Bhattacherjee, A., & Park, S. C. (2014). Why end-users move to the cloud: A migration-theoretic analysis. *European Journal of Information Systems, 23*(3), 357–372.
8. BongKeun, J., Kexin, Z., & Moutaz, K. (2012). Consumer piracy risk: Conceptualization and measurement in music sharing. *International Journal of Electronic Commerce, 16*(3), 89–118.
9. Boss, S. R., Galletta, D. F., Lowry, P. B., Moody, G. D., & Polak, P. (2015). What do systems users have to fear? Using fear appeals to engender threats and fear that motivate protective security behaviors. *MIS Quarterly, 39*(4), 837–864.
10. Chatterjee, S., Sarker, S., & Valacich, J. S. (2015). The behavioral roots of information systems security: Exploring key factors related to unethical IT use. *Journal of Management Information Systems, 31*(4), 49–87.
11. Chen, Y., & Zahedi, F. M. (2016). Individuals' internet security perception and behaviors: polycontextual contrasts between the United States and China. *MIS Quarterly, 40*(1), 205–222.
12. Choi, B. C. F., Kim, S. S., & Jiang, Z. (2016). Influence of firm's recovery endeavors upon privacy breach on online customer behavior. *Journal of Management Information Systems, 33*(3), 904–933.
13. Choudhary, V., & Zhang, Z. (2015). Patching the cloud: The impact of SaaS on patching strategy and the timing of software release. *Information Systems Research, 26*(4), 845–858.
14. D'Arcy, J., Herath, T., & Shoss, M. K. (2014). Understanding employee responses to stressful information security requirements: A coping perspective. *Journal of Management Information Systems, 31*(2), 285–318.
15. Forrest, C. (2016). Report: 80% of businesses can't properly manage external cyber attacks. Retrieved from http://www.techrepublic.com/article/report-80-of-businesses-cant-properly-manage-external-cyber-attacks
16. Foth, M. (2016). Factors influencing the intention to comply with data protection regulations in hospitals: Based on gender differences in behaviour and deterrence. *European Journal of Information Systems, 25*(2), 91–109.
17. Goode, S., Hoehle, H., Venkatesh, V., & Brown, S. A. (2017). User compensation as a data breach recovery action: An investigation of the sony playstation network breach. *MIS Quarterly, 41*(3), 703–728.
18. Guo, H., Cheng, H. K., & Kelley, K. (2016). Impact of network structure on malware propagation: A growth curve perspective. *Journal of Management Information Systems, 33*(1), 296–325.
19. Han, W. C., Ada, S., Sharman, R., & Rao, H. R. (2015). Campus emergency notification systems: An examination of factors affecting compliance with alerts. *MIS Quarterly, 39*(4), 909–930.

20. Herath, T., Chen, R., Wang, J. G., Banjara, K., Wilbur, J., & Rao, H. R. (2014). Security services as coping mechanisms: An investigation into user intention to adopt an email authentication service. *Information Systems Journal, 24*(1), 61–84.
21. Ho, S. M., Hancock, J. T., Booth, C., & Liu, X. W. (2016). Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication. *Journal of Management Information Systems, 33*(2), 393–420.
22. Hsu, J. S. C., Shih, S. P., Hung, Y. W., & Lowry, P. B. (2015). The role of extra-role behaviors and social controls in information security policy effectiveness. *Information Systems Research, 26*(2), 282–300.
23. Hu, Q., West, R., & Smarandescu, L. (2015). The Role of self-control in information security violations: Insights from a cognitive neuroscience perspective. *Journal of Management Information Systems, 31*(4), 6–48.
24. Hui, K.-L., Kim, S. H., & Wang, Q.-H. (2017). Cybercrime deterrence and international legislation: Evidence from distributed denial of service attacks. *MIS Quarterly, 41*(2), 497–524.
25. Jenkins, J. L., Anderson, B. B., Vance, A., Kirwan, C. B., & Eargle, D. (2016). More Harm Than Good? How messages that interrupt can make us vulnerable. *Information Systems Research, 27*(4), 880–896.
26. Jensen, M. L., Dinger, M., Wright, R. T., & Thatcher, J. B. (2017). Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems, 34*(2), 597–626.
27. Ji, Y., Kumar, S., & Mookerjee, V. (2016). When being hot is not cool: Monitoring hot lists for information security. *Information Systems Research, 27*(4), 897–918.
28. Johnston, A. C., Warkentin, M., McBride, M., & Carter, L. (2016). Dispositional and situational factors: Influences on information security policy violations. *European Journal of Information Systems, 25*(3), 231–251.
29. Johnston, A. C., Warkentin, M., & Siponen, M. (2015). An enhanced fear appeal rhetorical framework: Leveraging threats to the human asset through sanctioning rhetoric. *MIS Quarterly, 39*(1), 113–134.
30. Khansa, L., Kuem, J., Siponen, M., & Kim, S. S. (2017). To Cyberloaf or Not to Cyberloaf: The impact of the announcement of formal organizational controls. *Journal of Management Information Systems, 34*(1), 141–176.
31. Kim, D. J., Yim, M. S., Sugumaran, V., & Rao, H. R. (2016). Web assurance seal services, trust and consumers' concerns: an investigation of e-commerce transaction intentions across two nations. *European Journal of Information Systems, 25*(3), 252–273.
32. Kim, S. H., & Kim, B. C. (2014). Differential effects of prior experience on the malware resolution process. *MIS Quarterly, 38*(3), 655–678.
33. Kohli, R., & Tan, S. S.-L. (2016). Electronic health records: How can IS researchers contribute to transforming healthcare? *MIS Quarterly, 40*(3), 553–573.
34. Lee, D. J., Ahn, J. H., & Bang, Y. (2011). Managing consumer privacy concerns in personalization: a strategic analysis of privacy protection. *MIS Quarterly, 35*(2), 423–444.
35. Lee, C. H., Geng, X. J., & Raghunathan, S. (2016). Mandatory standards and organizational information security. *Information Systems Research, 27*(1), 70–86.
36. Li, H., Sarathy, R., Zhang, J., & Luo, X. (2014a). Exploring the effects of organizational justice, personal ethics and sanction on internet use policy compliance. *Information Systems Journal, 24*(6), 479–502
37. Li, L., Gao, P., & Mao, J-Y. (2014b). Research on IT in China: A call for greater contextualization. *Journal of Information Technology*, *29*, 208–222.
38. Li, W., Chen, H., & Nunamaker, J. F., Jr. (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems, 33*(4), 1059–1086.
39. Li, X. B., & Sarkar, S. (2014). Digression and value concatenation to enable privacy-preserving regression. *MIS Quarterly, 38*(3), 679–698.
40. Liang, N., Biros, D. P., & Luse, A. (2016). An empirical validation of malicious insider characteristics. *Journal of Management Information Systems, 33*(2), 361–392.

41. Lowry, P. B., & Moody, G. D. (2015). Proposing the control-reactance compliance model (CRCM) to explain opposing motivations to comply with organisational information security policies. *Information Systems Journal, 25*(5), 433–463.
42. Lowry, P. B., Posey, C., Bennett, R. J., & Roberts, T. L. (2015). Leveraging fairness and reactance theories to deter reactive computer abuse following enhanced organisational information security policies: An empirical study of the influence of counterfactual reasoning and organisational trust. *Information Systems Journal, 25*(3), 193–230.
43. Mitra, S., & Ransbotham, S. (2015). Information disclosure and the diffusion of information security attacks. *Information Systems Research, 26*(3), 565–584.
44. Ngai, E. W. T., & Wat, F. K. T. (2002). A literature review and classification of electronic commerce research. *Information & Management, 39*(5), 415–429.
45. Niemimaa, E., & Niemimaa, M. (2017). Information systems security policy implementation in practice: From best practices to situated practices. *European Journal of Information Systems, 26*(1), 1–20.
46. Oetzel, M. C., & Spiekermann, S. (2014). A systematic methodology for privacy impact assessments: A design science approach. *European Journal of Information Systems, 23*(2), 126–150.
47. Parks, R., Xu, H., Chu, C.-H., & Lowry, P. B. (2017). Examining the intended and unintended consequences of organisational privacy safeguards. *European Journal of Information Systems, 26*(1), 37–65.
48. Posey, C., Roberts, T. L., & Lowry, P. B. (2015). The impact of organizational commitment on insiders' motivation to protect organizational information assets. *Journal of Management Information Systems, 32*(4), 179–214.
49. Rossnagel, H., Zibuschka, J., Hinz, O., & Muntermann, J. (2014). Users' willingness to pay for web identity management systems. *European Journal of Information Systems, 23*(1), 36–50.
50. Sen, R., & Borle, S. (2015). Estimating the contextual risk of data breach: An empirical approach. *Journal of Management Information Systems, 32*(2), 314–341.
51. Siering, M., Koch, J.-A., & Deokar, A. V. (2016). Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems, 33*(2), 421–455.
52. Siponen, M., & Vance, A. (2014). Guidelines for improving the contextual relevance of field surveys: the case of information security policy violations. *European Journal of Information Systems, 23*(3), 289–305.
53. Sojer, M., Alexy, O., Kleinknecht, S., & Henkel, J. (2014). Understanding the drivers of unethical programming behavior: The inappropriate reuse of internet-accessible code. *Journal of Management Information Systems, 31*(3), 287–325.
54. Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdiscipliary review. *MIS Quarterly, 35*(4), 989–1015.
55. Steinbart, P. J., Keith, M. J., & Babb, J. (2016). Examining the continuance of secure behavior: A longitudinal field study of mobile device authentication. *Information Systems Research, 27*(2), 219–239.
56. Sutanto, J., Palme, E., Tan, C. H., & Phang, C. W. (2013). Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. *MIS Quarterly, 37*(4), 1141–1164.
57. Tsohou, A., Karyda, M., Kokolakis, S., & Kiountouzis, E. (2015). Managing the introduction of information security awareness programmes in organisations. *European Journal of Information Systems, 24*(1), 38–58.
58. Twyman, N. W., Lowry, P. B., Burgoon, J. K., & Nunamaker, J. F. (2014). Autonomous scientifically controlled screening systems for detecting information purposely concealed by individuals. *Journal of Management Information Systems, 31*(3), 106–137.
59. Vance, A., Anderson, B. B., Kirwan, C. B., & Eargle, D. (2014). Using measures of risk perception to predict information security behavior: Insights from electroencephalography (EEG). *Journal of the Association for Information Systems, 15*(10), 679–722.
60. Vance, A., Lowry, P. B., & Eggett, D. (2015). Increasing accountability through user-interface design artifacts: A new approach to addressing the problem of access-policy violations. *MIS Quarterly, 39*(2), 345–402.

61. Wall, J. D., Lowry, P. B., & Barlow, J. B. (2016). Organizational violations of externally governed privacy and security rules: Explaining and predicting selective violations under conditions of strain and excess. *Journal of the Association for Information Systems, 17*(1), 39–76.
62. Wang, J., Li, Y., & Rao, H. R. (2017). Coping responses in phishing detection: An investigation of antecedents and consequences. *Information Systems Research, 28*(2), 378–396.
63. Wang, J. G., Gupta, M., & Rao, H. R. (2015). Insider threats in a financial institution: Analysis of attack-proneness of information systems applications. *MIS Quarterly, 39*(1), 91–U491.
64. Wang, J. G., Xiao, N., & Rao, H. R. (2015). An exploration of risk characteristics of information security threats and related public information search behavior. *Information Systems Research, 26*(3), 619–633.
65. Warkentin, M., Walden, E., Johnston, A. C., & Straub, D. W. (2016). Neural correlates of protection motivation for secure IT behaviors: An fMRI examination. *Journal of the Association for Information Systems, 17*(3), 194–215.
66. Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii.
67. Willison, R., & Warkentin, M. (2013). Beyond deterrence: an expanded view of employee computer abuse. *MIS Quarterly, 37*(1), 1–20.
68. Wolff, J. (2016). Perverse effects in defense of computer systems: When more is less. *Journal of Management Information Systems, 33*(2), 597–620.
69. Wright, R. T., Jensen, M. L., Thatcher, J. B., Dinger, M., & Marett, K. (2014). Influence techniques in phishing attacks: An examination of vulnerability and resistance. *Information Systems Research, 25*(2), 385–400.
70. Zafar, H., & Clark, J. G. (2009). Current State of information security research in IS. *Communication of Association Information Systems, 24,* 557–596.
71. Zahedi, F. M., Abbasi, A., & Chen, Y. (2015). Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *Journal of the Association, 16*(6), 448–484.

Deploying a Scalable Data Science Environment Using Docker

Sergio Martín-Santana, Carlos J. Pérez-González, Marcos Colebrook, José L. Roda-García and Pedro González-Yanes

1 Introduction

The NIST Big Data Working Group (NBD-WG) [1, 2] provides a nice definition on the concept of Data Science:

> Data Science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing. It can also be understood as the activities happening in the processing layer of the system architecture, against data stored in the data layer, in order to extract knowledge from the raw data.

This definition implies a data life cycle, which is the set of processes that transform raw data into valuable and actionable knowledge, by means of principles, techniques and methods from many disciplines and domains (see Fig. 1) within the context of Big Data Engineering. For a brief introduction and a recent state-of-the-art on the concept of Big Data, the reader is referred to [3].

S. Martín-Santana
Máster en Ingeniería Informática, Universidad de La Laguna, Tenerife, Spain
e-mail: Sergio.MS.91@gmail.com

C. J. Pérez-González
Departamento de Matemáticas, Investigación Operativa y Computación,
Universidad de La Laguna, Tenerife, Spain
e-mail: cpgonzal@ull.edu.es

M. Colebrook (✉) · J. L. Roda-García
Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Tenerife, Spain
e-mail: mcolesan@ull.edu.es

J. L. Roda-García
e-mail: jlroda@ull.edu.es

P. González-Yanes
Centro de Cálculo de la Escuela Superior de Ingeniería y Tecnología (Secc. Ing. Informática),
Universidad de La Laguna, Tenerife, Spain
e-mail: pgonyan@ull.edu.es

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_7

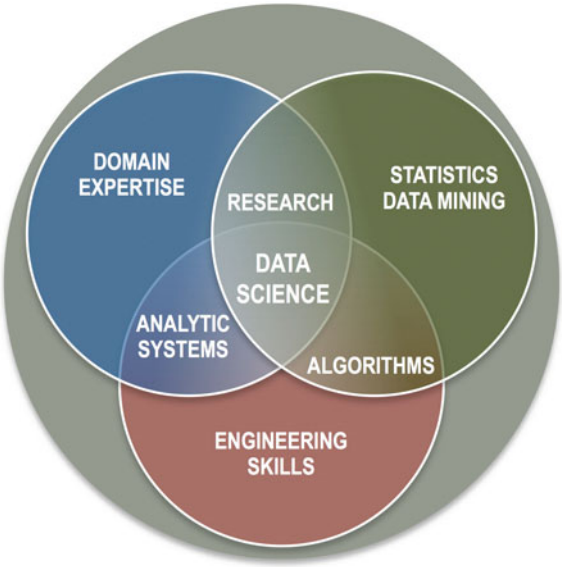


Fig. 1 Data Science definition from the point of view of the skills needed (adapted from [4])

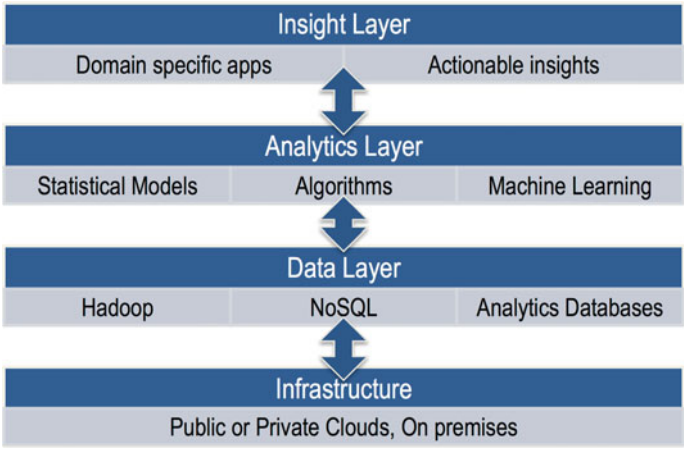


Fig. 2 Data science stack (adapted from [5])

Furthermore, such data life cycle is developed inside a Data Science stack (see Fig. 2), in which the infrastructure layer supporting the distributed computing engine plays an important role in order to obtain timely and accurate insights in a digital business.

Indeed, the market research company Forrester [6] pointed out that a new type of company has arisen nowadays: the **insights-driven business**, which builds sys-

tems using Data Science platforms to create competitive advantage through data. Moreover, it is predicted that these companies will earn a revenue of $1.2 trillion in 2020.

In this sense, companies that adopt a Data Driven Decision Making (DDDM) achieve a 5–6% increase in productivity and production growth [7]. Besides, the relationship between DDDM and performance appears also in other performance indicators such as asset utilization, return on equity and market value.

According to [8] there is no alternative: 65% of the firms think that there is a high risk of becoming uncompetitive if they do not implement a data driven mindset, since data is becoming a key component of their market value.

Forrester [6] also suggests that Data Science platforms, which comprise data integration, data exploration, model development and deployment, could accelerate insights maturity if the firms follow some key recommendations:

- Unify the Data Science technology into a single platform.
- Treat Data Science platforms as a strategic and transformative investment.

Within this enterprise context, Linden et al. from the consulting firm Gartner [9] define a Data Science platform as:

> A cohesive software application that offers a mixture of basic building blocks essential for creating all kinds of data science solutions, and for incorporating those solutions into business processes, surrounding infrastructure and products.

Additionally, their analysis of the 16 top vendors in Data Science platforms yields the following conclusions:

- The implementation of open source platforms is increasing the adoption of Data Science.
- Apache Spark is becoming a de facto Data Science foundation for the vendors.
- Open source languages like Python, R and Scala dominate this market. Even more, almost all Data Science platform vendors support Python and R.

Therefore, in order to facilitate the adoption of Data Science platforms, the Big Data Senior Steering Group (BD-SSG) [10] suggests to enhance infrastructures to support handling and analyzing large amounts of data, since state-of-the-art infrastructures are essential in a data-driven industry sector. They also noticed that there is a need to invest in infrastructure pilot programs, testbeds, and sandboxes for testing new techniques at scale, across a variety of application domains, and to engage in proofs of concept with both open source and proprietary solutions. Thus, future infrastructures may help moving the computation to the data.

Besides, the Big Data Value Association [11] also recommends building good infrastructures to develop a Data Economy, raising as a challenge a distributed trust infrastructure for data management, with flexible structures based on data microservices in a decentralized way. Regarding this matter, the European Union [12] is currently working in the development of enabling technologies, infrastructures and skills for the benefit of the SME (Small and Medium-sized Enterprises).

Likewise, the Edison Data Science Framework [2] promotes infrastructures, including typical frameworks such as Hadoop [13] and Spark [14], to support data handling during the whole data lifecycle. On the other hand, the NBD-WG [15] suggests creating a vendor-neutral, technology- and infrastructure-independent framework that could enable stakeholders using the best analytics tools on the most suitable computing platform and cluster. Besides, in order to support Big Data stores and processing, the infrastructure should be scalable in terms of easy addition of new resources, with possible platforms including public and/or private clouds [1].

Nevertheless, digital businesses investing only in infrastructure projects are not guaranteed to succeed, as pointed out by the UK's Science and Technology Committee of the House of Commons [16]. Acquiring more digital skills, trusting on public data sharing, progressing in open data and data protection are essential factors to remain in the right pace for Big Data and Data Science. Furthermore, the UK's government has been committed to creating a coordinated infrastructure, and access to advanced software and hardware to the small businesses (SME).

From the above paragraphs, it is clearly stated that the infrastructure layer plays an outstanding role within the Data Science stack. However, sometimes the expense of using such Data Science facilities in a private and commercial cloud infrastructure is not affordable to a small business. Accordingly, in the next sections we present a Data Science computing environment based on open source software tools that can be easily deployed over commodity (personal) computers.

Finally, the remainder of the chapter is organized as follows. In Sect. 2, we show the most important tools and environments for Data Science nowadays. Section 3 presents the full project and simple guides on how to deploy our Data Science stack in Windows, Linux or Mac. This stack has been used to analyze data from meteorological stations located in the Canary Islands (Spain), and the results are presented in Sect. 4. Finally, the conclusions are provided in Sect. 5.

2 Tools and Frameworks for Data Science

In Data Science there are many tasks that must be carried out frequently. For instance, loading and processing datasets, obtaining summarized statistics, visualizing the information in tables and charts, etc. The amount of tools and applications that are available to accomplish these jobs has increased in the last years, which implies installing programs and libraries in desktop or server computers with all the problems derived of this process.

Among the main difficulties that usually arise are those concerning to errors due to not complying with the dependencies between the required software versions or the lack of experience of the users in dealing with these computer system aspects. In this sense, virtualization is the solution to afford these issues since it provides the possibility to create and deploy software-based systems (so called virtual machines and containers) that emulate the physical ones.

A virtual machine consists in a guest system that packages both the computer architecture and the software applications along with the operating systems (plus all the code and dependencies required) to be executed in the host system. A container represents another level of virtualization where the host operating system kernel and its resources are shared to allow the execution of multiple light-weight and isolated processes. Consequently, each container takes up less space than virtual machines (container images are typically moderate in size), and run almost instantly.

The containers technology helps setting up the collection of useful tools for different stages in a Data Science project. Thus, each container represents a recipe for each application that can be shared and versioned. In the following sections, we describe and discuss the most usual programming languages and developing frameworks in order to create the stack of containers for Data Science.

2.1 Containers in Data Science

Since its first appearance in 2013 [17], Docker containers have implied a big impact in simplifying the process to create Data Science stacks. Basically, containers are lightweight versions of traditional virtual machines but without the need of large amounts of storage space on servers (see Fig. 3). Besides, they can be easily created and deleted, and they boot up quickly. Restoring a normal virtual machine usually can imply excessive time to get going, but Docker containers start up almost immediately.

The containers run from images that are essentially snapshots of a running container at a particular time point. These images can be used as templates to create and run other containers. This is the main reason why they are important in Data Science, since images are created containing the required tools for doing data analysis, either for a general use or for specific analyses. Lots of base images of containers can be downloaded for free from registries like Docker Hub [18]. The key idea is that many

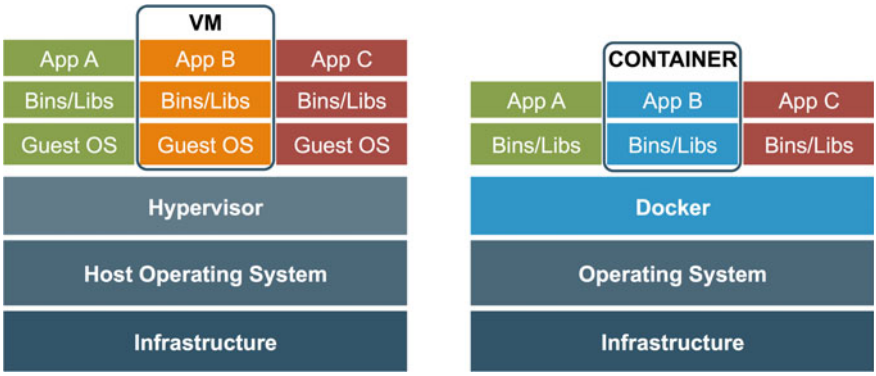


Fig. 3 Difference between virtual machines (*left*) and docker containers (*right*) (adapted from [17])

containers can be launched as required and, consequently, it turns into an easy task creating reproducible Data Science environments.

Running a container with the libraries and tools for a particular analysis reduces the effort to debug packages across different environments because they run identically on systems as Mac OS X, Windows or Linux. Due to this feature, Docker containers are very convenient to allow the users launching a variety of isolated applications in a platform as, for example, Jupyter [19] and RStudio [20] sessions configured with a set of basic packages, but also lending the users the possibility to install other libraries.

2.2 The R Language

The R language [21] represents the most well-known free, open source programming language and environment for statistical computing and graphics. Indeed, it is powerful and highly extensible with more than 10,000 add-on packages.

There are many large and active communities (for instance, the LinkedIn's R group has more than 100,000 members), and there are currently hundreds of R Meetup groups. This proves the increasing interest in the R statistics language, especially for data analysis. The programming environment allows for command-line scripting and, therefore, the data analysis steps can be serialized in such a way that can be reused with other data in contrast with interfaces guided with option menus.

The variety of tasks that can be accomplished in R are, among others, the following (we describe in parenthesis the aspects of data analysis that could be accomplished with these simple tasks):

- Exploring and manipulating data (ETL processing)
- Fitting and validation of predictive or classification models (machine learning)
- Creating visually attractive graphs (data visualization)
- Connecting with different data sources (systems integration)
- Making illustrative reports or dashboards (business intelligence).

The reader may find many R language tutorials in the Internet, some of them designed even for novice users without any programming background. These tutorials help users to understand the basics and fundamentals of R about importing and exporting data, exploring and manipulating data and, for advanced users, how to use loops and create functions.

R is one of the key tools in Data Science because it covers several data mining, machine learning and statistical techniques. There are also complete tutorials which explain how to perform descriptive statistics and make inferences on data, apply linear and logistic regression models as well as classification and clustering techniques, fit time series, apply variable selection and dimensionality reduction, etc.

2.3 RStudio

RStudio is an integrated development environment (IDE) that enhances the standard R and eases the work of R programmers [20]. It is available as open source for free, but there are also enterprise versions with additional features (administrative tools, enhanced security and authentication for multiple users, metrics and monitoring functionality, etc.).

RStudio is a very interesting application because it supports several premium characteristics such as intelligent code completion, syntax highlighting, integration of R help and the management of structured R documentation, and a tool for interactive debugging (see Fig. 4). The product can be used in a personal desktop installation or in a server version to centralize access and computation.

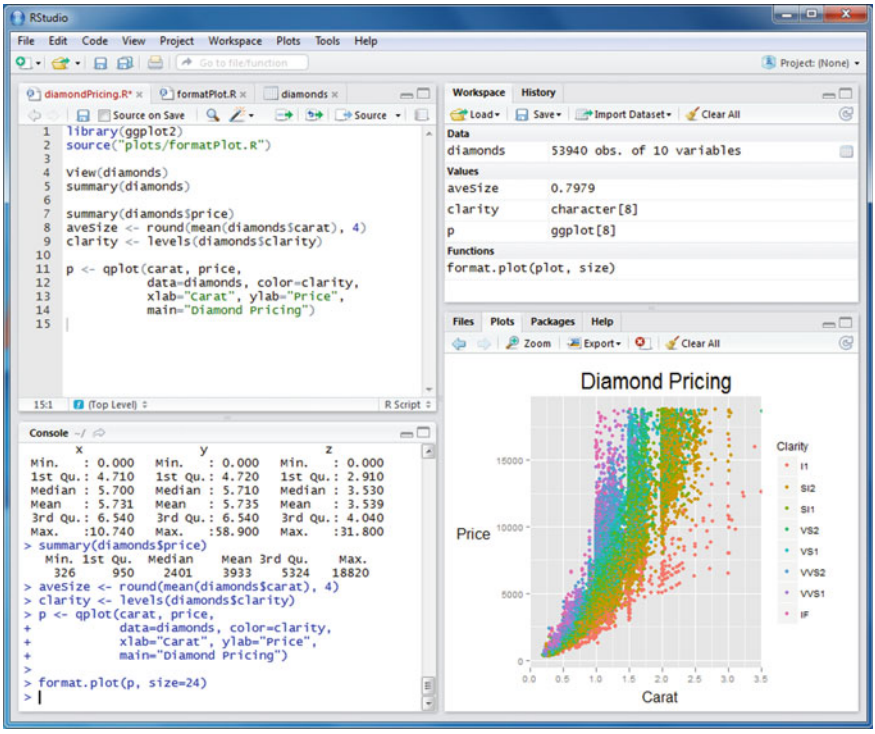


Fig. 4 The RStudio IDE (*source* [20])

2.4 The Python Language

Python [22] is a general purpose programming language and one of the most popular tools for data analysis. It is very frequent to use it when analyzing huge amounts of data due to several strengths. In a similar way to R, Python provides many powerful libraries appropriate to process very large and growing data sets, and there is a wide support from open source community users.

It is relatively easy to write code in Python and to make this code understandable by other users. Python also integrates very well with other open source platforms commonly used in Data Science, as Spark [14] and Hadoop [13]. These are the reasons that have contributed to the enthusiastic adoption of Python by the programmers.

A Python environment can be easily set up. There are free distributions like Anaconda [23] or Canopy [24] containing the core Python language, as well as other essential libraries for data analysis including the following:

- Numpy and Scipy: fundamental scientific computing
- Pandas: data manipulation and analysis
- Matplotlib: plotting and visualization
- Scikit-learn: machine learning and data mining
- StatsModels: statistical modeling, testing, and analysis.

Again as in the R case, there are many excellent internet resources (among others, DataCamp [25] and Codecademy [26]) to learn how to code in Python. They are an excellent option to gain knowledge in programming concepts that will be useful and valuable in working with data.

2.5 Jupyter Notebooks

One of the most important Python extensions is the Jupyter notebook (also known as IPython notebook) [19]. The notebooks are executable documents that, when launched from the Jupyter web interface, a browser is opened to show an environment to place not only code and executing data analysis, but even to introduce rich text, formatted expressions and embedded images and videos.

With Jupyter is possible to include several kernels that are computational engines for executing code of many other languages apart from Python (as for example R). The notebooks also provide options to export the content in several formats including PDF, HTML and Markdown. Consequently, notebook documents can be used as reports containing both the analysis description and the final results (figures, tables and graphics).

Other interesting Python IDE for data analysis is Rodeo [27], from the Yhat company. This program is similar to RStudio for R, and can be seen as a simple, lightweight alternative front-end to the Python notebooks (Fig. 5).

The R and Python languages described above, as well as the RStudio and Jupyter development environments (IDE), are included within the applications layer, whereas

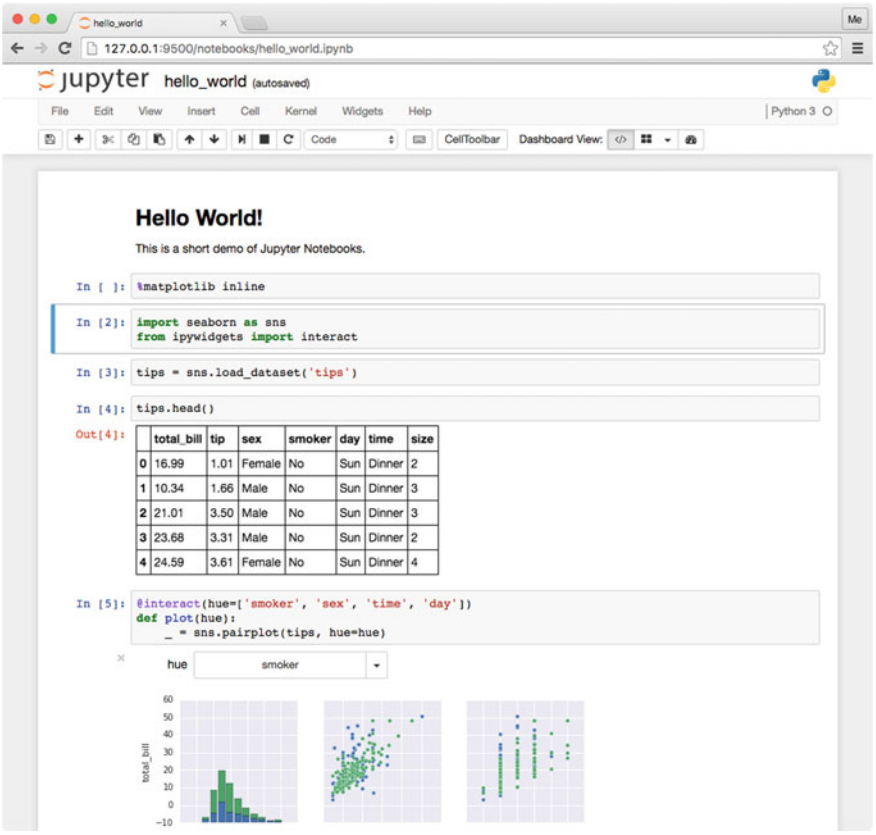


Fig. 5 An example of the Jupyter notebook (*source* [19])

the Docker containers are used in the infrastructure layer to ease the deployment process (see Fig. 2). The reader might have noticed that the data layer is missing in the previous schema. Such layer can be connected from inside the user container, as we explain in the following section where we describe the development of the Data Science Stack project.

3 The Data Science Stack Project

The integration of different languages, libraries and platforms for use in a real environment is a complex task. It is really crucial to model the ecosystem in which you are going to work. In this section, we describe the starting situation of this project, as well as the reasons to justify the change of the model towards a more efficient one.

We rely on the TOGAF architecture framework [28] that allows us to model, through the points of view, the existing system at the beginning of the project called AS-IS and after a set of change recommendations, model the target system called TO-BE. The Layered Viewpoint diagrams designed with the Archimate tool [29] will be used to describe both situations, and they are composed of three main layers: business, application and technology.

The key idea is to reflect in a single diagram the complete system. Besides, the Layered Viewpoint diagrams offer a joint vision of the actors, the main existing business processes, the software components and the technological infrastructure.

3.1 Initial Situation

Currently, there exists an infrastructure where the researcher users request the creation of different systems for data storage, data analysis and information visualization. One way to accomplish these tasks is through virtual machines that are generated from templates by system administrators. In this case, it is necessary to reserve the appropriate computing resources. This step involves several problems for the end user like waiting until the resources are available or the administration staff can complete the work.

To understand the starting point of our project, we will use a TOGAF-based diagram, which visually represents the AS-IS model [28]. This model is divided in layers where actors, processes, components and infrastructure are presented (the aforementioned business, application and technology layers).

In the current model, the execution for a research user of our system is presented, who may ask for a new infrastructure or simply use it in the system. To do so, this request should be solved by a system administrator. The application will be registered and, once accepted, it will be created. Then, the system can be really deployed. The business processes are executed in an application for the management of the forms and will have a mechanism to create virtual machines for later deployment. Finally, the complete system is executed according to the infrastructure layer presented, in which a web and virtualization servers are available (see Fig. 6).

3.2 Evolution of the Model to a More Efficient Solution

In order to speed up the deployment and implementation of Data Science systems, we propose a different strategy based in the dockerization of the system. Under this strategy, computer resources will be optimized, since no system resources reservation will be needed.

The proposed model aims to minimize and simplify the Administrator's task. Using the system described below, we could meet the requirements of the researchers without adding complex tasks. Initially, this solution will have two development

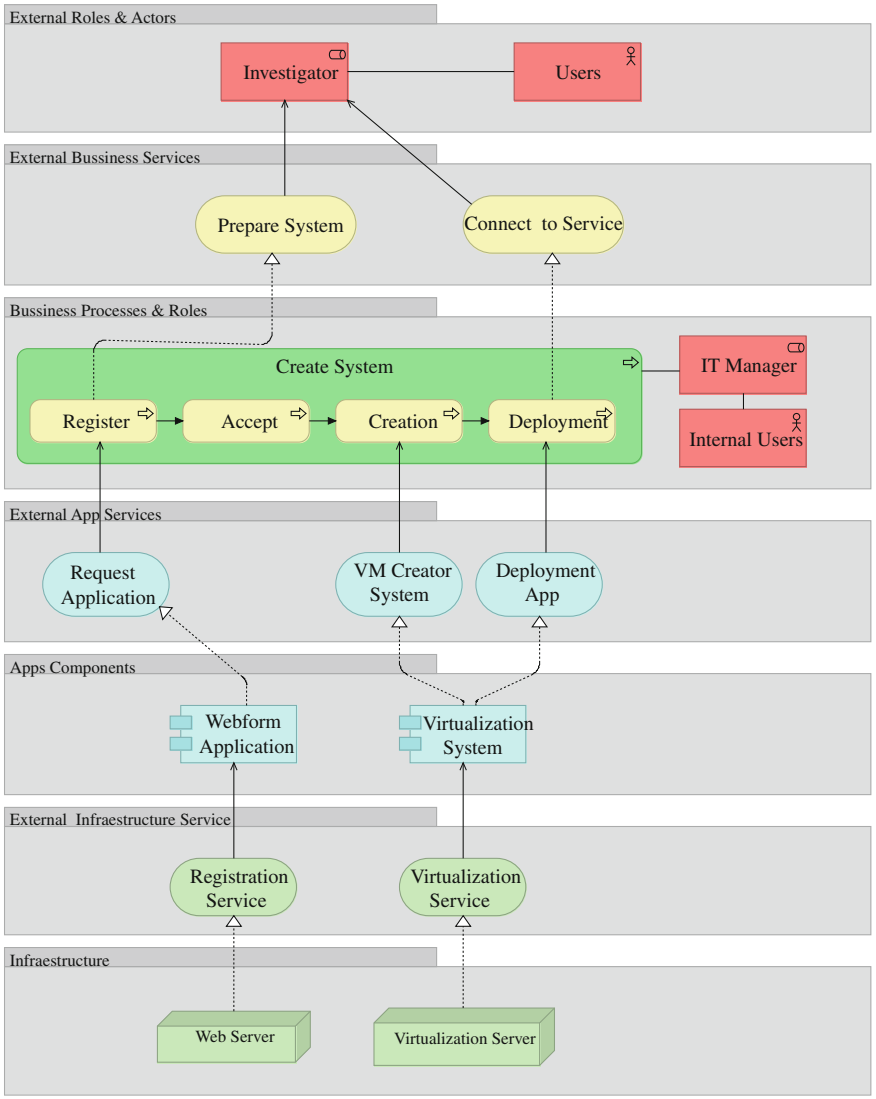


Fig. 6 The situation of the current system described as a TOGAF AS-IS diagram

environments, such as Jupyter and RStudio, which allow the applications development through languages such as Python and R. Besides, Spark is considered for the massive processing of data, and it also provides real-time data processing. All of these components run on a single Docker container, making a lot easier its maintenance and future evolution.

Hence, from the system administrator's point of view, his/her work could focus on creating a single, equal based recipe for all researchers that would include all

the required tools. From the point of view of the data scientist, he/she would have a complete set of tools for researching, with low cost extensibility to add new libraries, languages or tools.

This new solution avoids, or at least diminishes, the high costs of the equipment to allocate test systems. Now the researcher will be able to carry out small tests on his personal equipment, and to later transfer them to a production system in a large infrastructure.

The final solution is presented in Fig. 7 through a TOGAF-based model where the architecture of the system is visually displayed.

3.3 Solution Development: A Container Based in Docker

We have developed an environment with different applications installed as modules by executing a Dockerfile, which consists of several scripts doing specific tasks. Our proposal of this system uses the latest version of Ubuntu Xenial OS, but other operating systems can be used as Debian, CentOS, Alpine, etc., although this might imply some modifications of the steps described in the following paragraphs.

When the Dockerfile is executed, the first script (`base.sh`) installs some additional base system libraries and applications (e.g. `wget`, `default-jre`, `tar`, `openssl`, `libssl`, etc.) that are required for our Data Science environment.

In the next step of the Dockerfile execution, the scripts to install and configure the different applications of the Data Science stack are placed inside the container. This aspect provides modularity to our system because each tool is aggregated in an easy and independent way.

Thus, the first applications to be installed are versions 2.7 and 3.6 of the Python programming language using the instructions contained in the script named `python.sh`.

In addition to Python, the next scripts (`R.sh` and `Rstudio.sh`) install the last version of the R language, along with the integrated development environment (IDE) RStudio server. Both applications are configured to place the packages libraries into our system project folder and define the web server ports to access the programs by using a simple web browser (Chrome, Firefox, Safari, etc.).

The installation of the Spark framework for distributed computing is carried out by a script called `spark.sh`. This application can be used only with one master node (standalone mode), but the execution in cluster mode (with additional worker nodes) is possible by configuring some environment variables defined in the script `SparkConf.sh`. This script allows the users to modify several characteristics of the worker nodes as the number of CPUs assigned to them or the RAM available for each one.

Another important development tool that we aggregated to the Data Science stack is Jupyter, which provides the functionality to use both Python and R in the notebooks. The script `jupyter.sh` installs the main program and the kernel libraries to allow

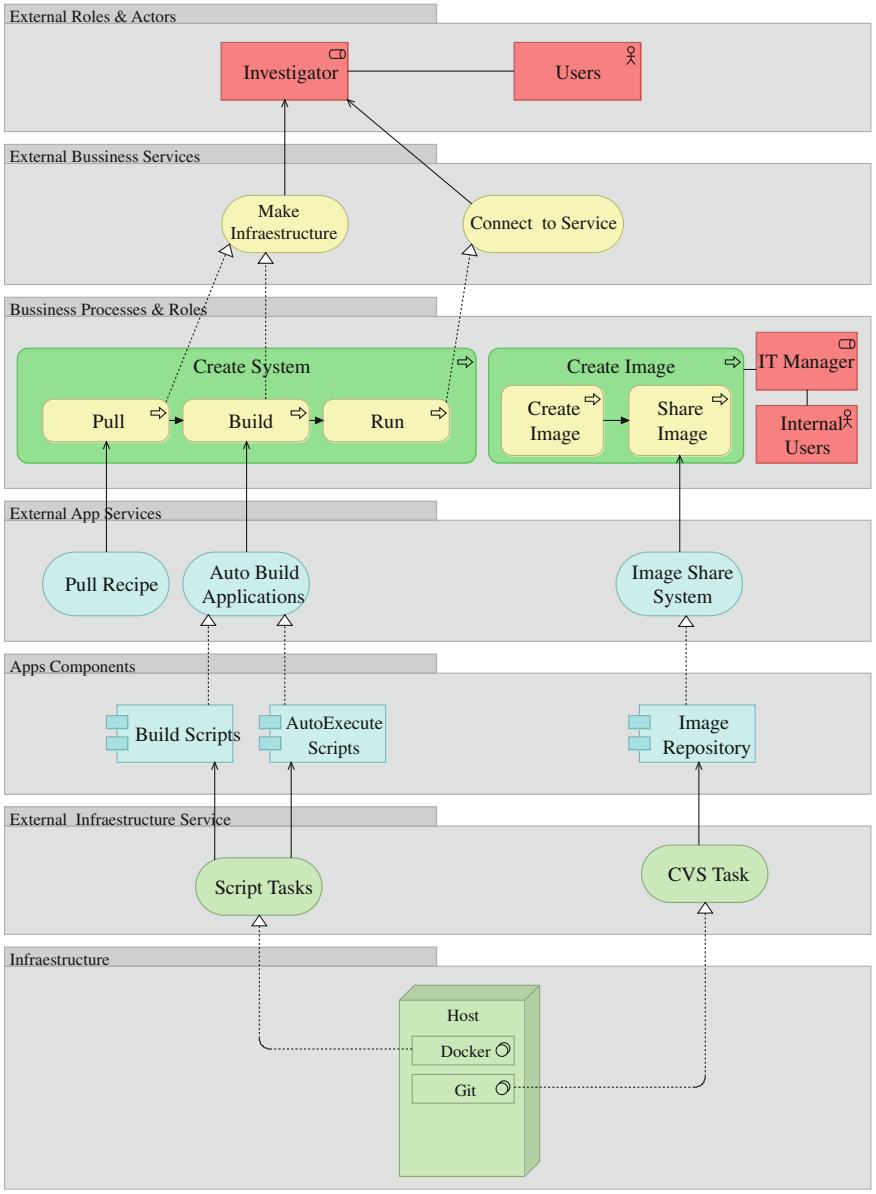


Fig. 7 The TOGAF TO-BE diagram as an evolution of the TOGAF AS-IS diagram

the execution of those languages in a notebook, as well as the configuration of the port to access the application through the web browser.

Table 1 Web applications with default and configured ports in the data science stack and links for accessing them

Application	Default port	Configured ports in deployment	Address link of application
RStudio	8787	10087	http://localhost:10087 (user: rstudio, password: rstudio)
Spark	8080, 8081, 7077	10080	http://localhost:10080
Jupyter	8888	10088	http://localhost:10088

Table 1 shows the list of the applications configured in the Data Science stack with their corresponding ports and addresses to be run using a web browser. Besides, Table 2 summarizes the list of scripts and files used in the deployment.

After the Dockerfile execution, the user can access the different applications by typing the links of Table 1 in the address bar of any browser, so now the system is ready to be used for the required Data Science tasks.

3.4 Deployment of the System

In the procedure to launch the environment, two scripts called `start_master` and `start_workers` are provided. The first one creates the network to run the container or cluster, and creates the user folder to place the notebooks in the host machine of the system. This folder is mounted as a data volume of the container and allows the users to access the files of the data stack environment from the host computer where this system is running. Besides, several R and Python libraries will be installed only the first time this script is run.

The second script is optional and enables the user to configure a cluster of N additional system instances to be run as workers. This cluster is valid for using the Spark framework, where the number of instances can be introduced in the `NWorkers` argument of the script.

In the following paragraphs, we describe the steps to deploy the data stack environment in a host computer with Windows and Linux Ubuntu operating systems.

Install and Run in Windows (64 Bits)

1. Open a command prompt (CMD) or Powershell using the Windows menu options.
2. Download the Github project typing:

```
git clone https://github.com/taroull/DockerForScience.git
--config core.autocrlf=input
```

Table 2 A summary of all the files used in the deployment

Located in	File name	Function
/	build_image.bat	Builds the container image in Windows
	build_image.sh	Builds the container image in Ubuntu or Mac
	Dockerfile	Docker instructions to build the images
install/	base.sh	Installs base system libraries and applications
	custom_python.sh	Allows installing certain versions of additional python libraries
	jupyter.sh	Installs Jupyter notebook and languages' kernels
	jupyter_notebook_config.py	Jupyter notebook configuration file
	PyLibraries.sh	Basic Python libraries
	python.sh	Installs versions 2.7 and 3.6 of Python
	R.sh	Installs the R programming language
	Rconfig.R	Installs the R kernel in Jupyter
	RStudio.sh	Installs the RStudio IDE
	spark.sh	Installs Spark
	SparkConf.sh	Spark configuration file
	start.sh	Entry point in Dockerfile: runs the container
Start/Ubuntu_Mac/	Docker_install.sh	Installs Docker in Ubuntu or Mac
	execute_worker.sh	Deploys one Spark worker node
	start_master.sh	Starts the Spark master
	start_workers.sh	Deploys additional worker nodes for Spark
Start/Windows/	runworker.bat	Deploys one Spark worker node
	start_master.bat	Starts the Spark master
	start_workers.bat	Deploys additional worker nodes for Spark

3. Change to working directory typing: `cd DockerForScience`
4. If Docker is not installed yet, the user can download and install it from the official web page [17]. When Docker is installed, a process window is started and pinned into a quick access of the desktop. Before proceeding, users must enable the "Shared Drives" option in the settings.

> In some systems, the user might need to configure appropriately the firewall or antivirus program to avoid problems when executing the following steps.

5. Optionally, the script `install\custom_python.sh` might be edited to include in variable `Libraries` some additional python libraries (separated by spaces) that the user wants to install in the stack. For example:

```
#!/bin/bash
#Specify as "<library>[==version] … [<libraryN>[==version]]"
Libraries="pandas==0.21.0 scipy bokeh plotly"
```

6. Execute the script `.\build_image.bat`, placed in the root folder of the project, to generate the Docker image in the host system. This process may take some minutes depending on file sizes and the internet connection speed.
7. After creating the image, the system can be started by executing `Start\Windows\start_master.bat`.
8. Consequently, all applications in Table 1 are accessible by typing the corresponding address link within a web browser.
9. Optionally, we can use the script `Start\Windows\start_workers.bat <NWorkers>` to deploy additional worker nodes for Spark. The number of slaves is specified in `<NWorkers>`, being 2 the default value of this parameter.

Install and run in Ubuntu/Debian or Mac OSX

1. Open a linux terminal.
2. Download the Github project using:

```
git clone https://github.com/taroull/DockerForScience.git
--config core.autocrlf=input
```

3. Change to the working directory by typing: `cd DockerForScience`

4. If Docker is not installed already, the user can execute the install script `Start/Ubuntu_Mac/Docker_install.sh`
5. Optionally, the script `install/custom_python.sh` might be edited to include in variable `Libraries` some additional python libraries (separated by spaces) that the user wants to install in the Data Science stack. See the example shown in point 5 of the previous section.
6. Execute the script `./build_image.sh` (it may require preceding a `sudo` command if the user has no admin privileges), placed in the root folder of the project, to generate the Docker image in the host system. The building process may take some minutes depending on file sizes and the internet connection speed.
7. After creating the built image, the system can be started by executing `Start/Ubuntu_Mac/start_master.sh`.
8. Now the applications in Table 1 are accessible by typing the corresponding address link in a web browser.
9. Furthermore, and only for Spark, we can use the script `Start/Ubuntu_Mac/start_workers.sh <NWorkers>` to start additional worker nodes, where the number of instances is set through the parameter `NWorkers`. The default value of this parameter is two instances.

Finally, and regardless of the system considered (Windows, Ubuntu or Mac), the user might access the container through the following command:

```
docker exec -ti master bash
```

Bear in mind that the deployment process might take a while depending on the speed of both your computer and the internet connection. Once we have set up our Data Science stack, the three main applications shown in Table 1 are now running in their corresponding ports at `localhost`, and ready to be used in a web browser (see Figs. 8, 9 and 10). Therefore, we can now proceed to show a case study in the next section.

4 A Case Study

This section presents an illustrative example of data analysis by using the Data Science stack developed in this chapter. In particular, the problem consists in studying the hourly observations collected in meteorological stations located in the Canary Islands (Spain) in order to obtain different statistical results using a Jupyter notebook and several R libraries. These stations are included in the United States Air Force (USAF) Master Station Catalog and registered in the National Climatic Data Center's (NCDC) archive of weather and climate data.

The notebook can be downloaded by running the following command in the notebooks folder:

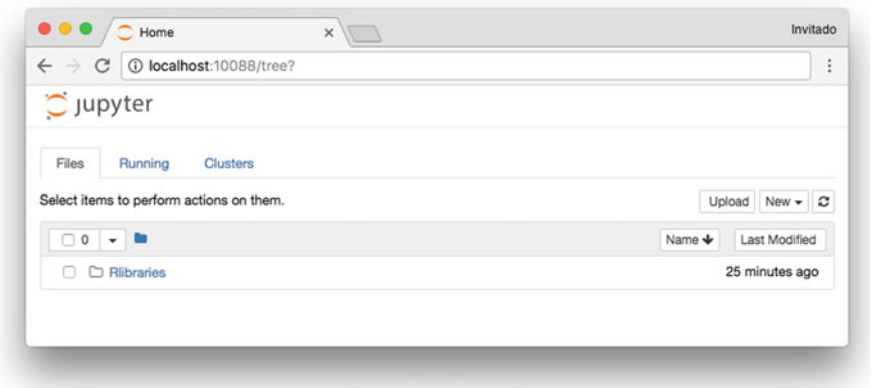


Fig. 8 The Jupyter notebook running in a web browser at localhost:10088

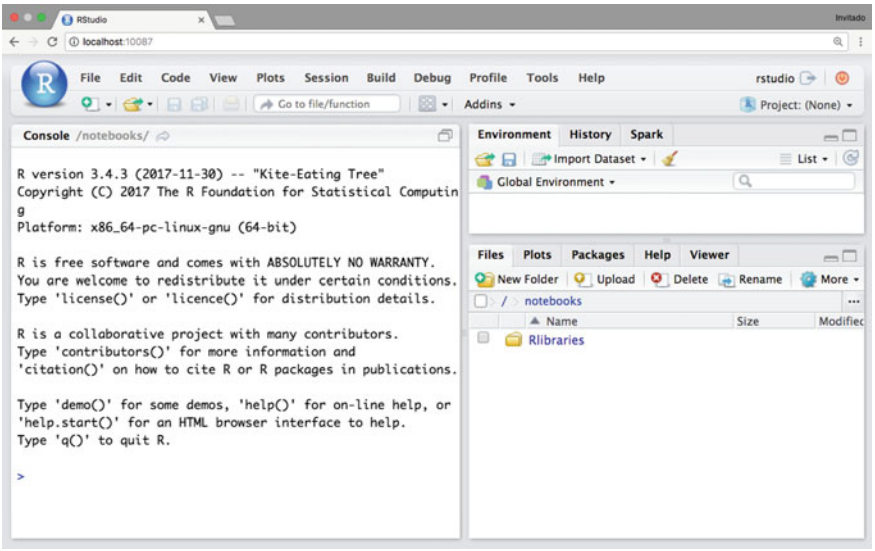


Fig. 9 RStudio running in a web browser

```
git clone https://github.com/taroull/Notebook-DataSciencewithR
```

For a given station, we can collect information by providing the beginning and ending years. Particularly, using the appropriate function in R, we get a data frame with information on hourly meteorological observations specifying a geographical bounding box and/or time bounds.

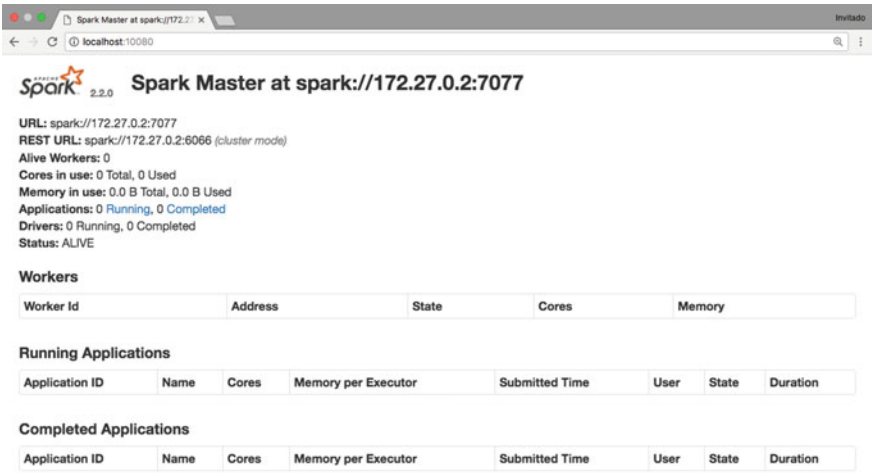


Fig. 10 The Spark Master deployed and running at localhost:10080

```
stations_canary$offset<-0.04
stations_canary$offset[stations_canary$name %in% c("IZANA","TENERIFE NORTE","TENERIFE SUR")]<-0.05

mapPoints <- ggmap(map) +
    geom_point(aes(x = lon, y = lat), data = stations_canary, col="red", alpha = .5) +
    geom_text(aes(x = lon, y = lat+offset, label=name), data = stations_canary, size = 2.0, hjust=
"left")+
    xlab("Longitude (degrees)") + ylab("Latitude (degrees)")
print(mapPoints)
```

Fig. 11 R script to draw the stations in a map

In this sense, the geographical position of the stations can be represented by using the `ggmap` library in R. The subsequent map (see Fig. 12) is easily obtained (see Fig. 11), and it can give us a good reference of the main locations where the data processed in the study were collected.

In particular, we are interested in studying the weather conditions in 2017 in the Teide National Park, located in the island of Tenerife. Consequently, we get hourly information of the nearest station about time, wind speed and direction, and temperatures through 2017 by using the script described in Fig. 13, and the result is shown just underneath as a formatted table.

In order to analyze these data, first we can plot all these values (see Fig. 14) to represent the evolution of daily temperatures in the National Park in 2017 using `ggplot`. The plot (see Fig. 15) uses a color gradient to appreciate the transition from the cold temperatures (blue) to the warm ones (red), and includes a smooth regression model fitting to the data.

Another interesting result consists in representing the distribution of low and high temperatures along the different daily hours in 2017. The chart (see Fig. 17) allows us to make a comparison study of which parts of the day are colder or warmer. Again,

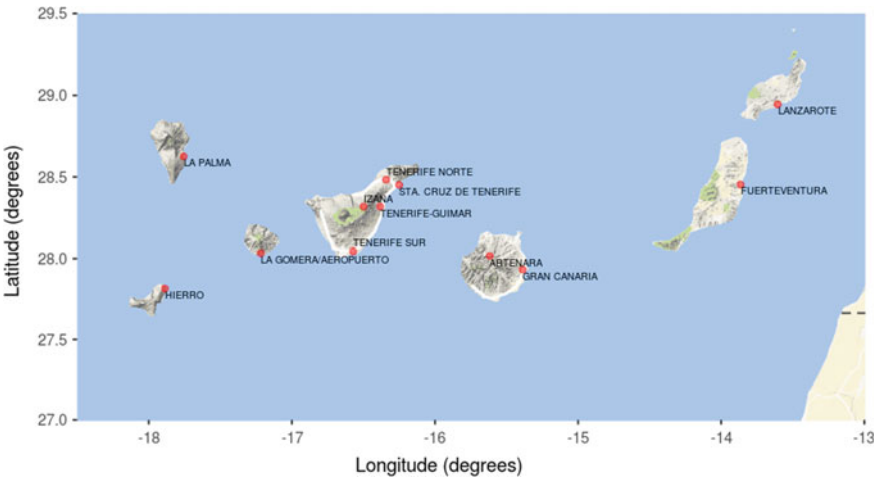


Fig. 12 The meteorological stations drawn on a map of the Canary Islands

```
temps_at_izana$time<-as.POSIXct(strptime(temps_at_izana$time,format="%Y-%m-%d %H:%M:%S"),format="%
Y-%m-%d %H:%M:%S")
head(temps_at_izana)
```

time	wind_direction	wind_speed	temp
2017-01-01 22:00:00	260	11.3	5.0
2017-01-01 23:00:00	250	10.8	4.5
2017-01-02 00:00:00	260	11.3	5.5
2017-01-02 01:00:00	260	10.8	5.7
2017-01-02 02:00:00	250	12.9	5.1
2017-01-02 03:00:00	250	13.4	4.2

Fig. 13 The meteorological data of the Teide National Park's station

```
ggplot(temps_at_izana,aes(x = time,y = temp)) +
  geom_point(aes(colour = temp)) +
  scale_colour_gradient2(name = "Temperature level",low = "blue", mid = "green" , high = "red", mi
dpoint = 12) +
  geom_smooth(color = "red",size = 1) +
  scale_y_continuous(limits = c(-5,30), breaks = seq(-5,30,5)) +
  ggtitle ("Daily temperatures and local smooth fitting in 2017 \n Izaña (Canary Islands)") +
  xlab("Month") +  ylab ("Temperature ( ºC )") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`geom_smooth()` using method = 'gam'
```

Fig. 14 R script that plots the data of the Teide National Park

we can observe that this result is relatively simple to obtain with the `ggplot` library (see Fig. 16).

Lastly, if the volume or complexity of the data is too large to be processed in a normal way using R, the user could easily connect to a Spark cluster in order to run the required analysis and get the final insights.

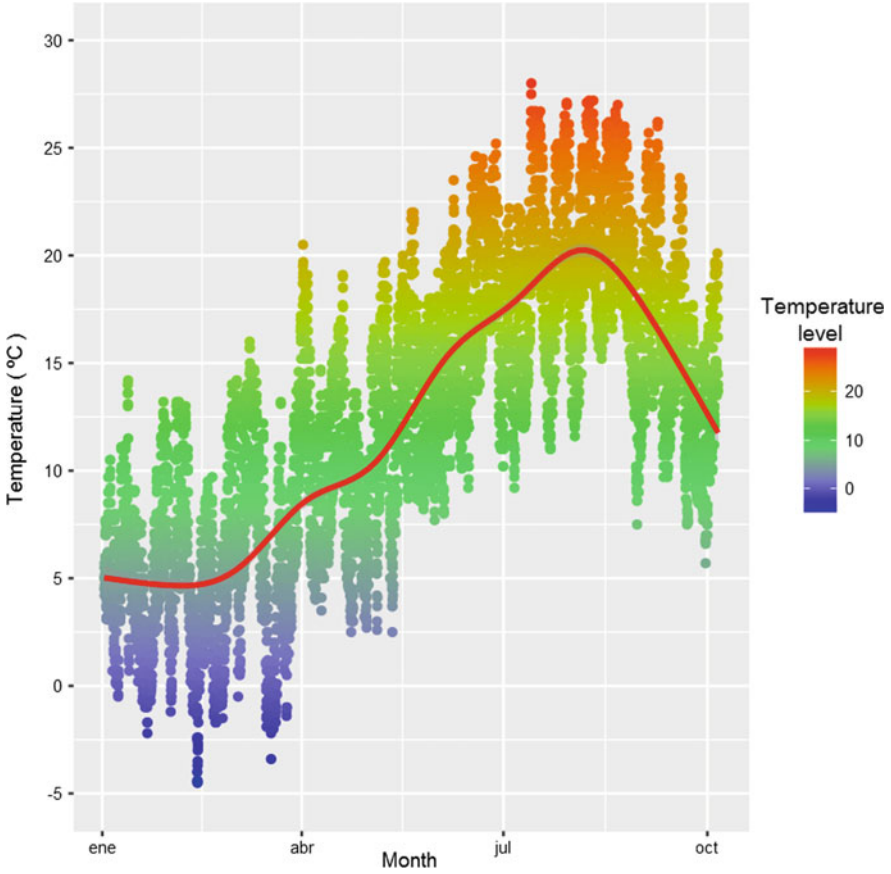


Fig. 15 Daily temperatures and local smooth fitting in Teide National Park in 2017

```
ggplot(temps_at_izana_temperatures) +
  geom_bar(aes(x = hour, fill = min.max.temp)) +
  scale_fill_discrete(name= "", labels = c("Daily high temperature","Daily low temperature")) +
  scale_y_continuous(limits = c(0,100)) +
  ggtitle ("Daily low and high temperatures distribution by hour \n Izaña (Canary Islands)") +
  xlab("Hour of the day") +  ylab ("Frequency") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Fig. 16 R script to plot the histogram of temperatures by hour

In this sense, and following the same example used in this section, we could now connect our notebook to a Spark master node to run some basic analysis (see Fig. 18).

Then, we copy our data to a Spark data frame to run some typical lazy Spark operations. In this case, we calculate the average temperatures grouped by month, as shown in Fig. 19.

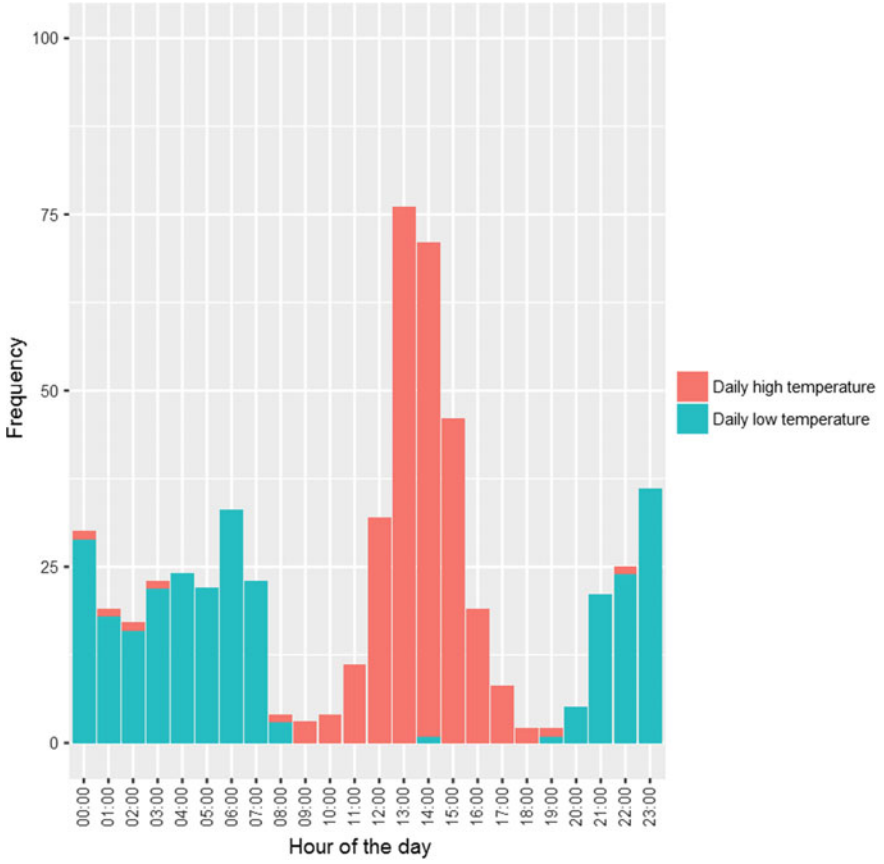


Fig. 17 Daily low and high temperature distribution by hour

```
usePackage("sparklyr")
```

```
Loading required package: sparklyr
```

```
conf <- spark_config()   # Load variable with spark_config()
conf[["sparklyr.defaultPackages"]] <- NULL
conf$spark.executor.memory <- "16G"
sc <- spark_connect(master = "local", spark_home="/usr/local/spark",
                    config = conf)
```

Fig. 18 R script to connect to a Spark master node

Finally, we use the usual `collect()` function in Spark to run query and return the results back to R (see Fig. 20).

The notebook is completed with further interesting results that provide a brief overview of the collected information. As a result, the notebook allows publishing not only the analysis but even the R script along with the data. This turns to be extremely useful in case other researchers want to reproduce and validate the study

```
temps_at_izana_avg<-temps_at_izana %>%
  group_by(month,month_name) %>%
  summarize(avg_wind=mean(wind_speed),avg_temp=mean(temp))
temps_at_izana_avg
```

```
# Source:   lazy query [?? x 4]
# Database: spark_connection
# Groups:   month
  month month_name avg_wind  avg_temp
  <chr>      <chr>    <dbl>     <dbl>
1    01      enero 9.073950  4.790476
2    02    febrero 7.803506  4.460518
3    03      marzo 7.712921  6.617900
4    04      abril 8.182260  9.586409
5    05       mayo 7.343103 12.443966
6    06      junio 5.969795 16.686657
7    07      julio 5.105517 18.588828
8    08     agosto 6.231077 19.353729
9    09 septiembre 7.713445 15.890588
```

Fig. 19 Example of some Spark lazy operations

```
temps_at_izana_avg <- collect(temps_at_izana_avg)
temps_at_izana_avg
```

month	month_name	avg_wind	avg_temp
01	enero	9.073950	4.790476
02	febrero	7.803506	4.460518
03	marzo	7.712921	6.617900
04	abril	8.182260	9.586409
05	mayo	7.343103	12.443966
06	junio	5.969795	16.686657
07	julio	5.105517	18.588828
08	agosto	6.231077	19.353729
09	septiembre	7.713445	15.890588

Fig. 20 Returning the average result from Spark back to R

without the complexity of installing all the required tools individually. Therefore, containers help to develop and deploy these complex setups in a simple way and run them in any environment.

5 Conclusions

By virtue of what has been stated in previous sections of this chapter, it has been clearly established that the infrastructure layer plays a key role within the Data Science stack. Indeed, there is a great concern on building good infrastructures to allow the stakeholders to run testbeds, sandboxes and proofs of concept (see [10–12, 15]).

On the other hand, a new type of company defined as the insights-driven business [6], that uses Data Science to create competitive advantage has arisen in the last years. Furthermore, companies that make data driven decisions can raise up to a 5–6% their productivity [7], or become completely uncompetitive otherwise [8].

However, some of these businesses cannot afford such Data Science services in a commercial cloud. Accordingly, and following the recommendations indicated in [9], we have developed a Data Science platform that can be easily deployed over commodity computers using open source software that includes Spark as the current de facto standard, as well as the R and Python languages.

To summarize, the main advantages of our Data Science stack approach are the following:

- The R libraries and the Python packages that are installed within the notebooks folder after the Docker image is created, can be reused when the container or the image are run again.
- The scripts can update in the future the applications already installed by just changing the variables defined inside them.
- The shared volume inside the Docker container allows the user saving the data and related files in a folder inside the host computer.
- As a result, the notebook allows publishing not only the analysis but even the R script along with the data. This turns to be extremely useful in case other researchers want to reproduce and validate the study.

Likewise, from the point of view of the main actors the advantages are:

- System administrators take less time to configure.
- Data analysts or researchers can really focus to accomplish their main duties rather than wasting time in administration tasks.
- The academia can benefit from the easiness to develop executable examples in a very simple way.

Besides, the project presented in this chapter can be very helpful for teaching and researching activities, and has been directly used in subjects regarding Laws and Regulations, Computer Vision, and Bioinformatics, and in a project on Earth and Atmosphere Observation Research. In this sense, the Data Science stack provides the students with the necessary tools for accomplishing different tasks in order to solve problems concerning Data Science (preprocessing data, statistical analysis, etc.), and preventing all the problems that usually arise when installing and configuring software.

In conclusion, our key objective has been essentially twofold. On the one hand, we have developed a simple, easy and fast platform to deploy a scalable Data Science stack that includes the main foundation tools. On the other hand, it has been designed for use in insights-driven businesses, as well as for obtaining reproducible results in research, and for teaching academic subjects easier.

Finally, the lines of future research are mainly focused in improving the current stack by including new Data Science environments like Zeppelin [30], and installing Deep Learning applications such as TensorFlow [31].

Acknowledgements This work is partially supported by the Spanish Ministry of Education and Science, Research Projects MTM2016-74877-P and CGL2015-67508-R, National Plan of Scientific

Research, Technological Development and Innovation. The authors wish to thank Adrián Muñoz-Barrera and Luis A. Rubio-Rodríguez for their support and assistance both in the configuration and deployment of the cluster and in the development of the solution.

References

1. NIST. (2015a). Big data interoperability framework: Volume 5, architectures white paper survey. Retrieved October 2017, from http://dx.doi.org/10.6028/NIST.SP.1500-5.
2. EDSF. (2017). The EDISON data science framework, Release 2. Retrieved October 2017, from http://edison-project.eu/edison/edison-data-science-framework-edsf.
3. Plaza-Martín, V., Pérez-González, C.J., Colebrook, M., Roda-García, J.L., González-Dos-Santos, T., & González-González, J.C. (2016). Analyzing network log files using big data techniques. In: F.P. García-Márquez, B. Lev (Eds.) *Big data management* (pp. 227–256). Springer International Publishing.
4. NIST. (2015b). Big data interoperability framework: Volume 1, definitions. Retrieved October 2017, from http://dx.doi.org/10.6028/NIST.SP.1500-1.
5. Hazard, C. (2014). Stacking the deck: The next wave of opportunity in big data. Retrieved October 2017, from https://www.kdnuggets.com/2014/05/stacking-deck-next-wave-opportunity-big-data.html.
6. Forrester. (2016). Data Science Platforms Help Companies Turn Data Into Business Value. Retrieved October 2017, from https://www.datascience.com/resources/white-papers/forrester-data-science-platforms.
7. Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decision making affect firm performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486.
8. Capgemini Consulting. (2015). Big & fast data: The rise of insight-driven business. Retrieved October 2017, from http://ww.capgemini.com/wp-content/uploads/2017/07/big_fast_data_the_rise_of_insight-driven_business-report.pdf.
9. Linden, A., Krensky, P., Hare, J., Idoine, C.J., Sicular, S., & Vashisth, S. (2017). magic quadrant for data science platforms. Retrieved October 2017, from https://www.gartner.com/doc/reprints?id=1-3TK9NW2&ct=170215&st=sb.
10. NITRD. (2016). The federal big data research and development strategic plan. Retrieved October 2017, from http://ww.nitrd.gov/PUBS/bigdatardstrategicplan.pdf.
11. BDV. (2017). Big data value strategic research and innovation agenda. Retrieved October 2017, from http://ww.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v3_0.pdf.
12. COTEC. (2017). Generación de talento Big Data en España (in Spanish). Retrieved October 2017, from http://cotec.es/media/BIG-DATA-FINAL-web.pdf.
13. Apache Hadoop. Retrieved October 2017, from http://hadoop.apache.org.
14. Apache Spark. Retrieved October 2017, from https://spark.apache.org.
15. NIST. (2015c). Big data interoperability framework: Volume 3, use cases and general requirements. Retrieved October 2017, from http://dx.doi.org/10.6028/NIST.SP.1500-3.
16. HC-STC. (2016). The big bata dilemma. Retrieved October 2017, from http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf.
17. Docker: The container platform provider. Retrieved October 2017, from http://www.docker.com.
18. Docker hub. Retrieved October 2017, from https://hub.docker.com.
19. Project Jupyter. Retrieved October 2017, from http://jupyter.org.
20. RStudio: The open source and enterprise-ready professional software for R. Retrieved October 2017, from https://www.rstudio.com.

21. The R Project for statistical computing. Retrieved October 2017, from https://www.r-project.org.
22. Python. Retrieved October 2017, from https://www.python.org.
23. Anaconda Python distribution. Retrieved October 2017, from https://www.anaconda.com/download/.
24. Enthought Canopy Python distribution. Retrieved October 2017, from https://www.enthought.com/product/canopy/.
25. Datacamp: Learn Data Science Online. Retrieved October 2017, from https://www.datacamp.com.
26. Codecademy: Learn to code interactively for free. Retrieved October 2017, from https://www.codecademy.com.
27. Rodeo: A Python IDE built for analyzing data. Retrieved October 2017, from https://www.datascience.com/blog/docker-containers-for-data-science.
28. The Open Group Architecture Framework (TOGAF) Version 9.1. The Open Group. Retrieved October 2017, from http://www.opengroup.org/togaf.
29. Lankhorst, M. M. (2004). Enterprise architecture modelling—the issue of integration. *Advanced Engineering Informatics, 18*(4), 205–216.
30. Zeppelin. Retrieved October 2017, from https://zeppelin.apache.org.
31. Tensorflow. Retrieved October 2017, from https://www.tensorflow.org.

Clean up CHAOS and Use E-CRM (A Digital Concept) to Expand the Business: A Case of Pakistan

Hina Amin and Abdullah Khan

1 Introduction

Technology is spreading its web to everywhere; survival of mankind depends on how promptly they respond or acquainted with the technology. Digital concept of CRM gives a great pathway to get the customer knowledge from a various sources and intelligently manage the database for required purposes. For making computer device sharper user should defragment it daily but it will done in an efficient way, alike that auto CRM system can be use to eliminate the chaos and manage the customers smoothly Here in chapter the concept of CRM is taken with the integration of technology as an E-CRM. An approach for multiple marketing channel technology is based on following points.

"Enterprise and scaled technology platform for various markets, these requires flexibility and desirability of globalization moreover tactical marketing capabilities which includes creativity with real time experimentation and optimization to all main touch points" [8]. In Pakistan, now customers are accepting gradually the concept of technology and using various aspects of technology in their daily routine whether it is a website or any app and at the same time organization also taking advantage of digitalization and get every information related to their customers/consumers easily from all digital platform including social media, web portal, smartphones etc. there are many solution based consultancies like Bitrix24 etc. that offers customize packages of e-CRM for organization and help to produce the customer pyramid that distinguish among potential, average and less profitable customers.

H. Amin · A. Khan (✉)
KASBIT, Karachi City, Pakistan
e-mail: Abdullah@kasbit.edu.pk

H. Amin
e-mail: hina.amin@kasbit.edu.pk

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_8

Another concept of digitalization has recently been announced that is known as "Bitcoin" a digital currency, electronic payment system proposed by one software developer to promote the exchange digitally it help to purchase the things electronically [3].

1.1 Background of CRM

The concept of CRM originated in early 1970s when the business had understanding that it would be better to become "Customer Emphatic" rather than focusing on products. The famous management guru and consultant Peter Drucker said that: "The real business of every company is to make and keep customers". CRM is comprises four dimensions marketing, sales, support and feedback of customers. CRM is fundamentally important for every business; it is about customer related process not just about technology [13].

1.2 Customer Relationship Management

> Always give people more than what they to get
>
> Nelson Boswell

Customer relationship management (CRM) receiving very much attention from the academic researchers and practitioners now a day why? Because of the philosophy that: "customer is everything" or "focus of marketing is to satisfy customers" [21].

As per the research more than $13 billion spent on CRM associated technology yearly worldwide [29]. CRM is actually an evolution of service marketing, in service marketing employees are more towards to satisfy customers through maintaining good relationship with them and enhance services to maximize retention [25]. Service marketing or CRM is based on two construct one is quality of services which are provided by manufacturer/service provider another is customer satisfaction this is the ultimate purpose of service marketing [28]. Another author defines CRM as "explaining, understanding and improving the marketing relationships in order to have the reputed survival" [6].

CRM is not just only about the latest technology it requires best integration of technology, process and efficient people for implementation [15].

1.3 Basic Framework of CRM

Here we have comprehensive diagram of CRM and its activities which generally practice in all countries (Fig. 1).

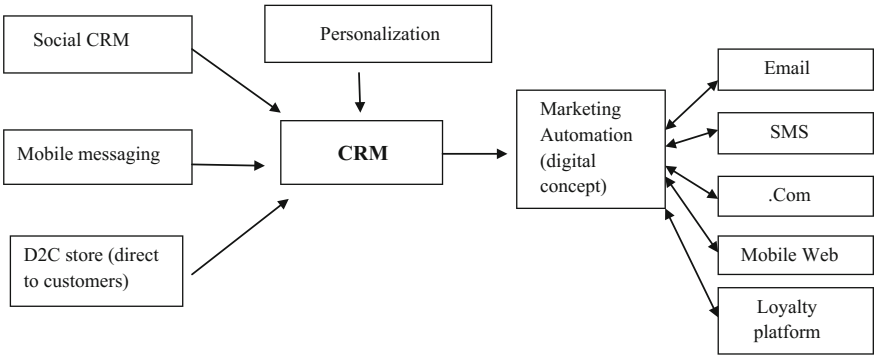


Fig. 1 Basic framework of CRM [8]

The above mentioned framework of CRM depicted the basic concepts of service marketing that includes various means to communicate with the customers and strategies to create good relationship with them for a long time period. The capabilities and strategies of CRM should be integrated not isolated for getting successful results.

To stay connected with the customers and looking for differentiated presence in the market, marketers should understand their CRM capabilities and try to exploit these capabilities in addition they should have customer profile or their 360° view that comprises demographical, social and shopping related data [8].

1.4 Components of CRM

The main issue with CRM is that it has its broad literature, for some managers CRM is to send emails for some others it is just to stay in touch with customers to increase their value and market worth. Here is a basic model of CRM given below which includes seven (7) important components which is quite essential from the marketing perspective (Fig. 2).

These components have identified by marketers and researchers, the improvements in innovative technologies have made easy to access with customers that ultimately results greater profitability [31], the bullet points of these components are:

- Extensive data base of customer activities.
- Proper database analysis.
- After getting the analysis results, decide about which customers to target.
- Take decisions about the tools for targeting customers.
- How to make relationship with the customers.
- Considers private issues occur in CRM.
- Create metric to assessing the success of company's CRM program and tools.

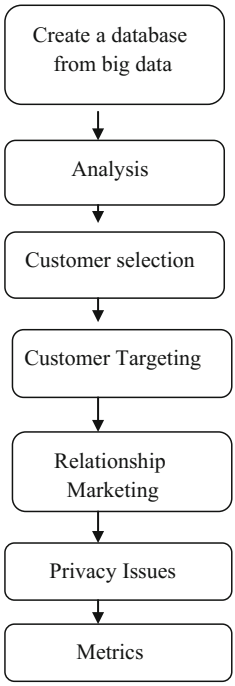


Fig. 2 Components of CRM [31]

Another purposeful viewpoint about CRM is that it consists of three (3) elements:

- Identifying, interacting, satisfying and maximizing the worth of organization key customers.
- Should have the knowledge of customer' needs and profile to ensure the contact with each customer is appropriate.
- Developing a complete picture of prospective customer [14].

1.5 Strategic Customer Relationship Management

Customer relationship management is highly recognized concept worldwide and it is considered as one of the important business initiative taken by most companies [22]. Research study identified that almost 1000 executives of international companies suggested: CRM is among 10 most important priorities in strategic moves for improving overall performance of organization (The Economist 2005).

CRM is not only the marketing process rather it is a core process of cross functional organizations it is not only enhance the shareholder value but also help to develop effective relationship with customers [17]. Organizational strategies play an important role for planning and implementing the successful CRM initiatives [5].

CRM helps to trigger the customers responsive in a way that helps to achieve the competitive advantage like this:

- Offer superior value to the customers and personalizing interaction.
- Show company's reliability and trustworthiness in front of the customers.
- Strengthen the relationship with the customers.
- Attains coordination in organizational capabilities.

CRM is much broader concept than just building a relationship with customers to fulfill their demands. To incorporate the CRM concept with technology and software is not sufficient it is basically a continuous process to make sound managerial decisions for the company through customer relationship with on individual basis [18].

It is also noticeable that CRM database and systems provide predictive analysis for instance, dissatisfaction of customers and rate of defects, etc.

1.6 E-CRM and Business Expansion

The main purpose of CRM application is to move from a product-focused to a customer-centric strategy this will lead to increase the value of the company and allow it to achieve competitive advantage over the rivals, because if customers will be satisfied with the performance of company then it will become easy to retain them in a right way [15]. CRM is defined as acquisition of knowledge about the customers. Social media revolutionized the way companies do their business and CRM, customer engagement is very much crucial now a day business should utilize the extensive information of the customer for making long term relation [2].

1.7 E-CRM in Pakistan

In banking sector of Pakistan the concept of CRM comprises certain dimension like resolution of customer's problem, right knowledge about customers and customer empowerment, this effort will be helpful to take competitive advantage and customer knowledge to develop strong CRM is very much vital [1], businesses are looking forward to opt every successful opportunity by applying relationship management tactics. Many businesses and marketing person in Pakistan argued that customer CRM helps to enhance the customer satisfaction and loyalty level among customers but one more is also noteworthy here is that marketer should understand the needs and requirements of their customers otherwise they will switch to another brand because they many choices in this turbulent market and role of escalating technology in the industry will also be an opportunity or threat depend on the usage of this technology with clears direction and purpose (winning maximum customers).

Khan et al. [12] identified that the main purpose behind investing in CRM is to automate the sales force and to provide the separate applications for CRM with their personal database. Many organizations in Pakistan have moved to data warehousing for using CRM application to maximize the benefits, now the question arises what is data warehousing, "Data warehousing is the facility that store the data centrally and collect the information from the various sources and provides access to many people (relevant stakeholders) to meet business intelligence and decision support requirements of organization". Consequently, companies are getting many advantages of CRM application with integration of warehouse usage like for instance timely and high quality data, reduction of operational cost, alignment with business goals, improved customer satisfaction and retention etc. [12]. In Pakistan customer are not ready yet to adopt the technology, which ultimately results as low level with e-banking that shows the lack of trust and low familiarity in technology [27].

1.8 Levels of CRM in Organization

Literature elucidates the levels of CRM that gives guidelines to implement it, CRM can be seen from: company-wide, Customer opinion and finally functional levels [13]. Company-wide perspective done on strategic level it provide the complete view, functional level consider the focus of that are require to fulfill the marketing activities while customer facing view offers the individual view of customers on this level marketers perform their duties on a continuing basis. However, distinction among following three areas of CRM is also considerable to know for better output; operational CRM that deals with interaction with customers, analytical CRM is about analysis of data about customers for prompt decision making and finally collaborative CRM which works for interdepartmental work and communication within organization to improve customer services and their experiences [7, 15].

1.9 Causes of Failure

There are many reasons of failure associated with CRM; this is due to underestimating crucial issues in implementing the customer relationship concepts in the organization, following are some points which show the causes of CRM failure:

- Before developing a customer strategy implementing CRM.
- Implement CRM concept before aligning the organizational strategies with it.
- Assume that CRM technology is better in every aspects.
- Investing more money on creating and maintaining relationship with impartial customers [24].

1.10 Successful Implementation of CRM

Sometimes due to improper implementation of CRM or failure of getting desire results many companies suffered from bad market reputation and damaged customer relationships. However, if they implement their CRM strategies carefully and efficiently they will get outstanding result [5].

Recommendations for successful implementation are given below:

- A front office that integrate all marketing function in media.
- A data warehouse that keeps all the relevant data and record.
- Development of proper business rules to get benefit through learning about the customers.
- For improvement purpose measures the performance of CRM continuously [17].
- Training your employees and always look for customers feedback.

When the CRM is aligned with electronic medium then it is called e-CRM, the implementation of e-CRM not only reduce the cost but also increase the profitability by putting a loyalty element in their strategies. Along with better e-service quality proper e-CRM execution of entire process and efforts of concern people will lead to positive word of mouth, retention, loyalty, repetition in purchase etc. [30].

1.11 CRM and Dynamic Capabilities

CRM and relationship marketing are almost similar concepts [9]. CRM is the narrower concept which only covers to build relationship with customers while relationship marketing focusing on whole range of stakeholders relationship [10].

Relationship with customers is the best way to achieve competitive advantage [4]. Dynamic capability in organization is extended theory of RBV (Resource based view) which emphasize on the internal strength of the firm and such strength should be performed in a superior way than competitors [10, 9]. This theory has applicable to CRM implementation in many large organizations [20]. Dynamic capabilities theory use to achieve performance benefits in a particular context of organization.

SOCIAL e-CRM (a new concept):

There are many customers active now on social media, for marketers social media become most powerful tool. The use of this tool by marketers has increasing day by day [10, 11]. Through the social media technology marketers and customers get close and interact by two ways. It is really very inexpensive, effective and marketers are able to get prompt feedback [16] and enable themselves to engage with customers.

1.12 Future of CRM (Digital Concept)

Increasing literature in CRM and in-depth philosophies given by marketers and by academic researchers give a rise to this phenomena and focusing on its proper implementation. It is crystal clear that the improvements happening in this area is marvelous companies to make long lasting and effective relationship with their customers, still markers is not in an ideal position. It is expecting that technologies and researches will further enhance the CRM discipline and its practicality. Companies are offering different incentives to the customers to give familiarity with their brands and seek higher level of loyalty from them [23].

One way to increase focus on CRM in organization is to create a separate managerial post or position for acquiring and retaining customers. Practicing CRM activities organization should perform cost and benefit analysis. Majority of the companies are getting benefits from the CRM activities perhaps only few companies found no such big difference in their market value and profitability [31]. CRM developed the unique identity in marketing discipline and it became the need of every business, the data of customer is instrumental to achieve organizational purpose because this data helps out to address the serving needs of customers [26], so this should be crucial to use the data of customers in a right way and consider as resource of the firm for value creation and new direction of data usage should be identified.

1.13 Recommendations

One of the biggest advantages of using CRM technology is that it manages all the contacts very easily without creating any chaos. Overlapping of tasks and responsibilities may cause the waste the times of employees and their energy. It is also observing that customers are not really satisfied with their contact experiences so they should train their employees to make customers delighted. The latest CRM concept is integrated with contact management. Industries are spending too much time for making their marketing efforts perfect. The future of CRM is to engage customers strongly, the most successful companies are those who customer centric so, marketers should become proactive and create ongoing interaction with customers to achieve higher profitability of the company. It is also important for organization to understand that value of core products in return of money play an important role in customer's satisfaction specifically in hotel industry so CRM can be enhance by providing better quality products to the respective customers [19].

References

1. Bhat, S. A., & Darzi, M. A. (2016). Customer relationship management: An approach to competitive advantage in the banking sector by exploring the mediational role of loyalty. *International Journal of Bank Marketing, 34*(3), 388–410.
2. Choudhury, M. M., & Harrigan, P. (2014). CRM to social CRM: The integration of new technologies into customer relationship management. *Journal of Strategic Marketing, 22*(2), 149–176.
3. Coindesk Inc. (2015). *What is bitcoin*. Retrieved from https://www.coindesk.com/information/what-is-bitcoin/.
4. Coltman, T. (2007). Why build a customer relationship management capability? *The Journal of Strategic Information Systems, 16,* 301–320.
5. Cravens, D. W., & Piercy, N. F. (2013). *Strategic marketing* (10th ed.). New York, USA: McGraw-Hill Companies Inc.
6. Dibb, S., & Meadows, M. (2004). Relationship marketing and CRM: A financial services case study. *Journal of Strategic Marketing, 12*(June), 111–125.
7. Fayerman, M. (2002). Customer relationship management. *New Dir ForInst Res* 113, 57–68.
8. Gupta. (2014). *Marketing technology: The need for Cadence around the chaos* (e–article), http://www.dmnews.com/marketing-strategy/marketing-technology-the-need-for-cadence-around-the-chaos/article/339621/.
9. Harker, M. J., & Egan, J. (2006). The past, present and future of relationship marketing. *Journal of Marketing Management, 22*(1–2), 215–242.
10. Harrigan, P., & Miles, M. (2014). From e-CRM to s-CRM. Critical factors underpinning the social CRM activities of SMEs. *Small Enterprise Research*, *21*(1), 99–116.
11. HubSpot (2013). *2013 State of inbound marketing annual report.* Retrieved November 2013, from http://offers.hubspot.com/2013-state-ofinbound-marketing?__hstc=20629287.d5b93ef1447e2227dfcbd1268526086e.1372405014143.1372405014143.1372405014143.1&__hssc=20629287.1.1372405014143.
12. Khan, A., Ehsan, N., Mirza, E., & Sarwar, S. Z. (2012). Integration between customer relationship management (CRM) and data warehousing. *Procedia Technology, 1,* 239–249.
13. Kumar, V. (2010). *Customer relationship management*. John Wiley & Sons, Ltd.
14. Maklan, S., Payne, A., Peppard, J., & Ryals, L. (2002). *Customer relationship management: Perspectives from the marketplace*. Taylor & Francis.
15. Orenga-Roglá, S., & Chalmeta, R. (2016). Social customer relationship management: Taking advantage of Web 2.0 and Big Data technologies. *SpringerPlus*, *5*(1), 1462.
16. Pagani, M., & Mirabello, A. (2012). The influence of personal and social-interactive engagement in social TV web sites. *International Journal of Electronic Commerce, 16*(2), 41–67.
17. Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing, 69*(4), 167–176.
18. Peppers, D., & Rogers, M. (1995). A new marketing paradigm. *Planning Review, 23*(2), 14–18.
19. Rahimi, R., & Kozak, M. (2017). Impact of customer relationship management on customer satisfaction: The case of a budget hotel chain. *Journal of Travel & Tourism Marketing, 34*(1), 40–51.
20. Rai, A., Patnayakuni, R., & Seth, N. (2006). Firm performance impacts of digitally enabled supply chain integration capabilities. *MIS quarterly*, 225–246.
21. Raman, P., Wittmann, C. M., & Rauseo, N. A. (2006). Leveraging CRM for sales: The role of organizational capabilities in successful CRM implementation. *Journal of Personal Selling & Sales Management, 26*(1), 39–53.
22. Ramaswami, S. N., Bhargava, M., & Srivastava, R. (2004). Market-based assets and capabilities, business processes, and financial performance. *MSI Reports, Marketing Science Institute* (*Hrsg.*), (04–102).
23. Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The customer relationship management process: Its measurement and impact on performance. *Journal of Marketing Research, 41*(3), 293–305.

24. Rigby, D. K., Reichheld, F. F., & Schefter, P. (2002). Avoid the four perils of CRM. *Harvard Business Review, 80*(2), 101–109.
25. Ryals, L., & Payne, A. (2001). Customer relationship management in financialservices: Towards information-enabled relationship marketing. *Journal of Strategic Marketing, 9,* 3–27.
26. Saarijärvi, H., Karjaluoto, H., & Kuusela, H. (2013). Customer relationship management: The evolving role of customer data. *Marketing Intelligence & Planning, 31*(6), 584–600.
27. Shakil Ahmad, M., Rashid, S., & Ehtisham-Ul-Mujeeb. (2012). ECRM and customers: A case of Askari Commercial Bank, Pakistan. *Business Strategy Series*, *13*(6), 323–330.
28. Taylor, S. A., & Baker, T. L. (1994). An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions. *Journal of Retailing, 70*(2), 163–178.
29. Thompson, B. (2003). What is CRM? customer think guide to real CRM, WhitePaper, CRMGuru.com, San Francisco. Retrieved from www.crmguru.com/member/papers/guide.pdf).
30. Wahab, S., Al-Momani, K., & Noor, N. A. M. (2015). The relationship between e-service quality and ease of use on customer relationship management (CRM) performance: An empirical investigation in Jordan mobile phone services. *The Journal of Internet Banking and Commerce 2010*.
31. Winer, R. S. (2001). A framework for customer relationship management. *California Management Review, 43*(4), 89–105.

Furthering Big Data Utilization in Tourism

Masahide Yamamoto

1 Introduction

In Japan, many people have struggled to promote tourism in their regions to vitalize their local economies. The low-cost carriers have boosted competition in the transportation industry, and domestic transportation costs have declined in several regions. Therefore, tourism is likely to become increasingly important to local economies. Initially, Japan's tourism industry suffered significant volatility in demand depending on the season and day of the week. In addition, there was significant loss of business opportunities because of congestion during the busy season.

To cope with such volatility, tourism facilities, such as inns and hotels, have been trying to level the demand through daily and/or seasonal pricing adjustments. For example, room rates on the days before holidays are usually more expensive than they are on other days. Despite these efforts, the differences between on-season and off-season occupancy rates of rooms and facilities are still large. In other words, attracting customers in the off-season is an important challenge for tourism. Various events have been held to eliminate the seasonal gap.

Numerous events are currently held to attract visitors in Japan. Many events are newly launched. To date, it has been difficult to accurately grasp the extent to which these events attract visitors and the types of people who visit. However, by employing the recently provided Information and Communication Technology (ICT) services, it is possible to verify the number and characteristics of visitors to a particular event.

In recent years, the so-called "*big data*" have been attracting the attention of companies and researchers worldwide. For example, convenience stores now can quickly predict sales of new products from the enormous amounts of information collected by cash register terminals and thereby optimize their purchases and inventories. Several governments and researchers have implemented such mechanisms in tourism.

M. Yamamoto (✉)
Faculty of Foreign Studies, Nagoya Gakuin University, Nagoya, Japan
e-mail: myama@ngu.ac.jp

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_9

In this chapter, I attempt to identify the number of visitors in different periods and their characteristics based on the location data of mobile phone users collected by the mobile phone company. In addition, I also attempt to demonstrate an alternative method to more accurately measure the number of visitors attracted by an event.

2 Attempts to Utilize Big Data in Tourism

So far, maximum attempts have been made in academic research to study the possibility of using big data in tourism. Previous tourism marketing research has primarily focused on the ways service promises are made and kept, mostly generating frameworks to improve managerial decisions or providing insights on associations between constructs [1]. *Big data* have become important in many research areas, such as data mining, machine learning, computational intelligence, information fusion, the semantic Web, and social networks [2]. To date, several attempts have been made to use large-scale data or mobile phone location data in tourism marketing studies.

Most studies dealing with big data in tourism were published after 2010. Fuchs et al. presented a knowledge infrastructure that has recently been implemented at the leading Swedish mountain tourism destination, Åre [3]. Using a Business Intelligence approach, the *Destination Management Information System Åre* (DMIS-Åre) drives knowledge creation and application as a precondition of organizational learning at tourism destinations.

Xiang et al. tried to apply big data to tourism marketing. The study aimed to explore and demonstrate the utility of big data analytics to better understand important hospitality issues, namely, the relationship between hotel guest experience and satisfaction [4]. Specifically, the investigators applied a text analytical approach to a large number of consumer reviews extracted from Expedia.com to deconstruct hotel guest experiences and examine the association with satisfaction ratings.

2.1 Utilization of Mobile Phone Users' Location Data in Tourism

Studies on using mobile phone location data for tourism surveys can be traced back to 2008. Ahas et al. introduced the applicability of passive mobile positioning data for studying tourism [5]. They used a database of roaming location (foreign phones) and call activities in network cells: the location, time, random identification, and country of origin of each called phone. Using examples from Estonia, their study described the peculiarities of the data, data gathering, sampling, the handling of the spatial database, and some analytical methods to demonstrate that mobile positioning data have valuable applications for geographic studies. Japan Tourism Agency conducted a similar study using international roaming service in December 2014 [6].

Since the creative work of Ahas et al. [5], several studies employing location data have emerged. Liu et al. investigated the extent to which behavioral routines could reveal the activities being performed at mobile phone call locations captured when users initiate or receive voice calls or messages [7]. Using data collected from the natural mobile phone communication patterns of 80 users over more than a year, they assessed the approach via a set of extensive experiments. Based on the ensemble of models, they achieved prediction accuracy of 69.7%. The experiment results demonstrated the potential to annotate mobile phone locations based on the integration of data mining techniques with the characteristics of underlying activity-travel behavior.

Alternative related studies have also been conducted. Gao and Liu attempted to examine the methods used to estimate traffic measures using information from mobile phones, accounting for the fact that each vehicle likely contains more than one phone because of the popularity of mobile phones [8]. Steenbruggen et al. used mobile phone data to provide new spatio-temporal tools for improving urban planning and reducing inefficiencies in current urban systems [9]. They addressed the applicability of such digital data to develop innovative applications to improve urban management.

As described above, I surveyed previous related research. Among those studies, this research could be characterized as similar to Ahas et al. [5]. However, Ahas et al. [5] is based on results obtained by analyzing data roaming activity. Mobile phone users in the study are obviously limited. Therefore, whether the knowledge gained applies to the average traveler is not clear. In this chapter, I analyze data provided by NTT DOCOMO, Inc., which is the largest mobile phone service provider in Japan. Therefore, their data should be more reliable in that the parameter is quite large.

Another attempt needs to be mentioned here. The Project Report that Okinawa Prefecture published is of a study that used location data obtained from a domestic mobile phone network [10]. The aim of the project was to survey the characteristics and behavior of tourists who were visiting Okinawa Prefecture. Okinawa pref. conducted the survey in order to grasp the trends and needs of repeat customers. The survey revealed the composition of tourists to Okinawa Prefecture by residence, gender, and age. They examined how the number of travelers changes depending on the month (October 2012 and January 2013) and the day of the week.

2.2 Mobile Kukan Toukei

In this chapter, I would like to introduce "*Mobile Kukan Toukei*™" provided by NTT DOCOMO, Inc. and DOCOMO Insight Marketing, Inc. I used the service to collect the location data of mobile phone users in order to count the number of visitors at specific tourist destinations and examine their characteristics (see Fig. 1).

Mobile Kukan Toukei is statistical population data created by a mobile phone network. It is possible to estimate the population structure of a region by gender, age, and residence using this service of a particular company. The locations and characteristics of the individuals obtained herein are derived through a non-identification process,

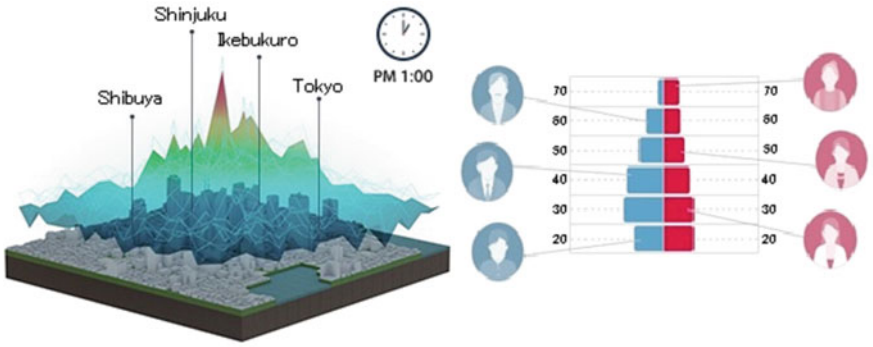


Fig. 1 Location data of mobile phone users collected from Mobile Kukan Toukei. Retrieved August 8, 2018, https://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/#p01

aggregation processing, and concealment processing. Therefore, it is impossible to identify specific individuals. The spatial statistical data from the mobile phone network were also used in the project on Okinawa Prefecture as mentioned above.

3 Utilizing Mobile Kukan Toukei to Examine the Effect of the Opening of the Shinkansen High-Speed Railway Line

The sites studied in this survey are tourist destinations in Ishikawa Prefecture and Toyama city, including Kanazawa city, which became nationally popular when the *Hokuriku Shinkansen* (high-speed railway) opened in 2015.

The survey areas are presented in Table 1 and Fig. 2. When selecting these areas, it was essential to identify their "regional mesh codes." A regional mesh code is a code for identifying the regional mesh. It stands for an encoded area that is substantially divided into the same size of a square (mesh) based on the latitude and longitude in order to use it for statistics. With regard to regional mesh, there are three types of meshes: primary, secondary, and tertiary. The length of one side of a primary mesh is about 80 km, and those of secondary and tertiary meshes are about 10 km and 1 km respectively.

In addition, split regional meshes also exist, which are a more detailed regional division. A half-regional mesh is a tertiary mesh that is divided into two equal pieces in the vertical and horizontal directions. The length of one side is about 500 m. Furthermore, the length of one side of a quarter and 1/8 regional meshes is about 250 m and 125 m respectively.

For example, the mesh code of Wakura Hot Springs, which is one of the survey areas, is a third order code 5536-5703 (Fig. 3). If the survey area cannot be covered in one mesh, it is possible to combine multiple meshes, like Kenrokuen in Table 1.

Table 1 Survey areas and regional mesh codes

	Survey areas	Regional mesh code	Type of codes
Kanazawa	Kanazawa Station	5436-6591-2	1/2
	Kenrokuen	5436-6572+5436-6573-1, 5436-6573-3	Tertiary, 1/2
	Higashi Chayagai	5436-6583-3	1/2
Nanao	Wakura Hot Springs	5536-5703	Tertiary
	Nanao Station	5536-4757	Tertiary
Kaga	Yamanaka Hot Springs	5436-2299, 5436-2390	Tertiary
Wajima	Wajima	5636-0772	Tertiary
Toyama	Toyama Station	5537-0147-1	1/2

Note A regional mesh code is a code for identifying the regional mesh, which is substantially divided into the same size of a square (mesh) based on the latitude and longitude in order to use it for statistics. The length of one side of a primary mesh is about 80 km, and those of secondary and tertiary meshes are about 10 km and 1 km respectively

4 Examining the Effect of the Opening of the Hokuriku Shinkansen

This study analyzed the location data collected from NTT DOCOMO, Inc. to consider the effect of the opening of the *Hokuriku Shinkansen* on the survey areas. The periods during which the data were collected are 8.00–9.00, 12.00–13.00, and 14.00–15.00 h.

4.1 Transition in Number of Visitors in Each Time Period

In general, the number of visitors has been increasing since the *Hokuriku Shinkansen* was launched on May 14, 2015, with the exception of Nanao station. It should be noted that these "visitors" also include the residents living there, because the data cannot exclude them. Of course, I tried to exclude residential areas as much as possible when I specified the regional mesh codes. However, it was rather difficult to do that, because the mesh codes are square-shaped.

4.1.1 Kanazawa City

First, I examined the data of Kanazawa city (see Figs. 4 and 5). There were two survey areas in this city: Kanazawa station and Kenrokuen Park.

Fig. 2 Survey areas

4.1.2 Toyama City

I then compared the results of two larger cities, Kanazawa and Toyama. Both these cities have a station at which the Hokuriku Shinkansen stops. It should be noted that Kanazawa city and Toyama attracted more visitors in the afternoons (see Fig. 6), whereas Wakura and Wajima, which are located on the Noto Peninsula, had more visitors in the mornings (8:00 a.m.–9:00 a.m.).

Visitors to Toyama Station demonstrated approximately the same trend as those visiting Kanazawa Station. However, there were fewer visitors on holidays than on weekdays in Toyama.

4.1.3 Wajima and Nanao City

Wajima is famous for its *Asaichi* (morning market). Presumably, visitors enjoy shopping at the Noto Shokusai market near Nanao Station during the daytime, move on

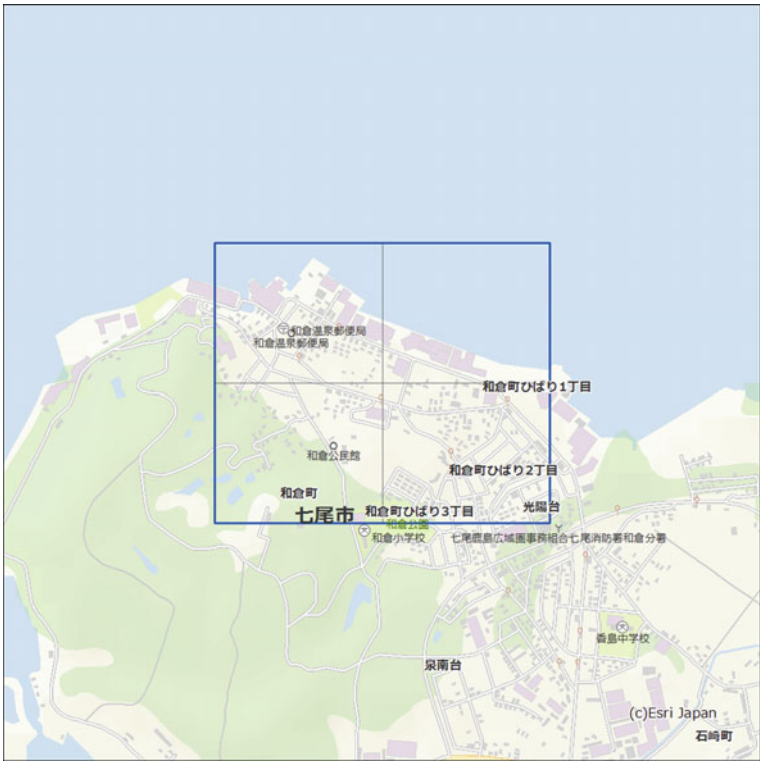


Fig. 3 The regional mesh of Wakura Hot Springs

to Wakura hot springs later in the day, and then return to the morning market the next day (Figs. 7 and 8).

Wajima is also the setting for a TV drama "Mare," which was broadcasted nationwide from April to September 2015. Visits to Wajima have slightly increased in 2015. The increase in visits could be attributed to this TV drama rather than the opening of the Hokuriku Shinkansen, because Nanao attracted fewer visitors in 2015 than in 2014 in spite of being a better location and nearer to Kanazawa.

4.1.4 Hot Springs

Regarding Yamanaka hot springs, there was no significant difference in visits between different periods. Some visitors might have spent more than one night in Yamanaka hot springs (see Fig. 9).

Although both Wakura and Yamanaka hot springs are a little far from Kanazawa, their results were contrary. The TV drama might have increased the number of tourists in Wakura (Fig. 10).

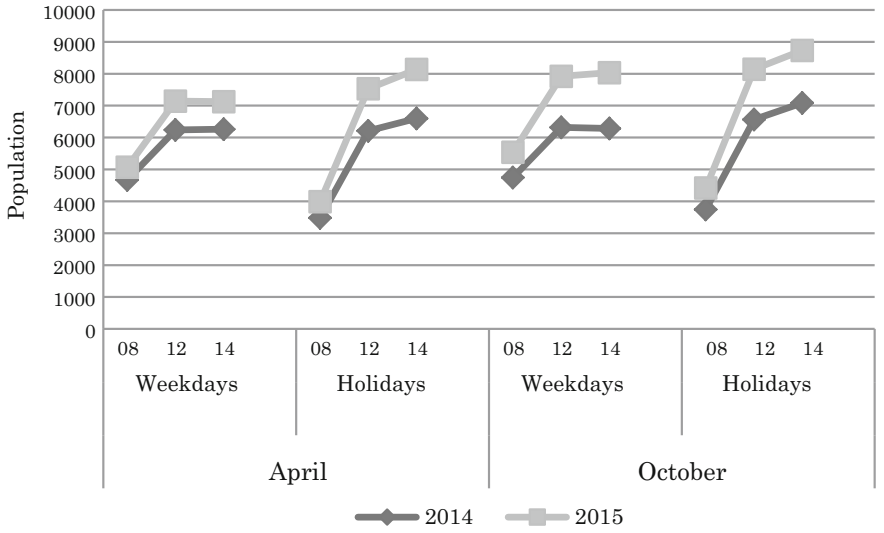


Fig. 4 Visitor transitions at Kanazawa Station

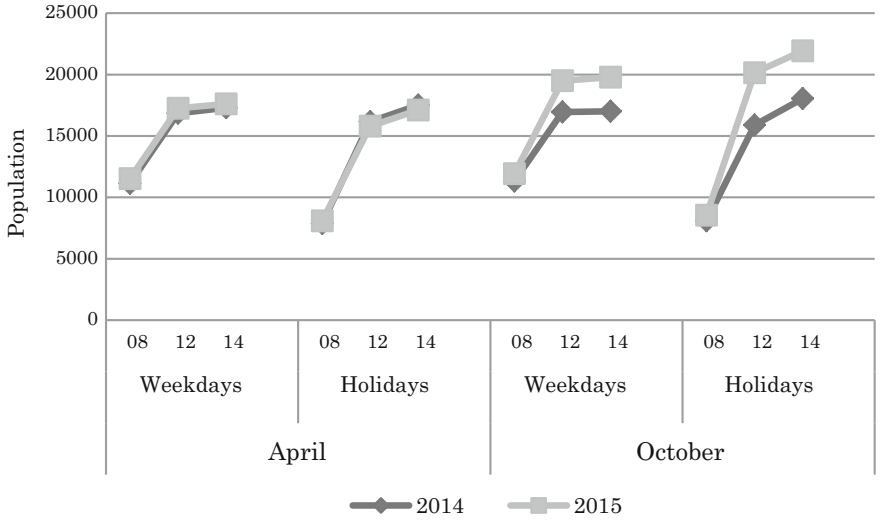


Fig. 5 Visitor transitions at Kenrokuen

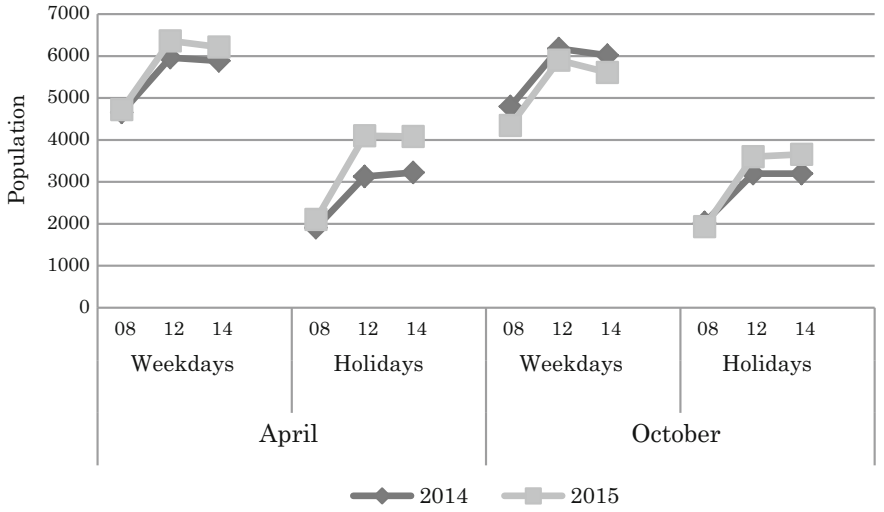


Fig. 6 Visitor transitions at Toyama Station

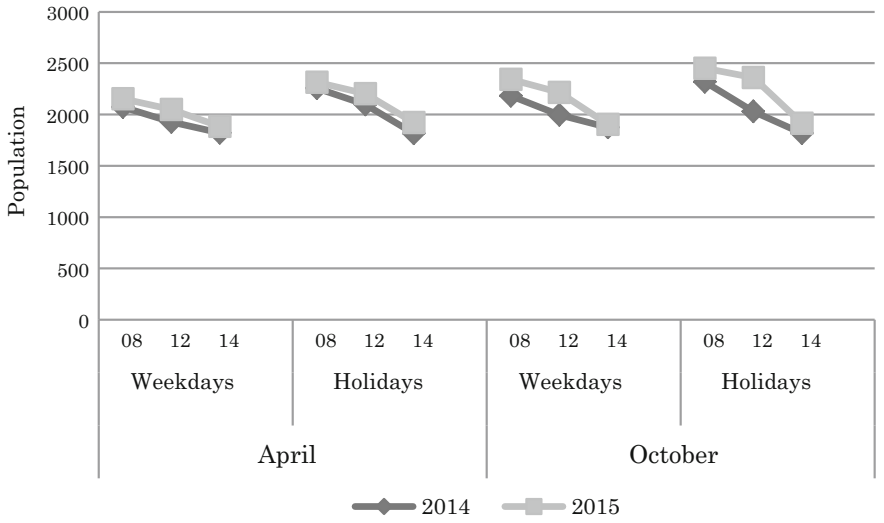


Fig. 7 Visitor transitions at Wajima

4.2 Visitors' Characteristics: Gender, Age, and Residence

4.2.1 Visitors' Gender and Age

Kanazawa city (particularly Kenrokuen) attracted a variety of visitors, including many female visitors (see Fig. 11).

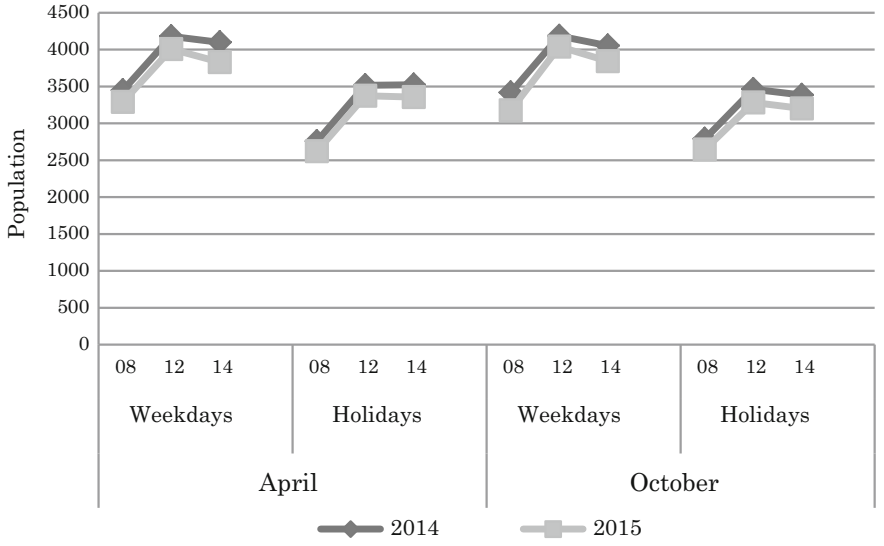


Fig. 8 Visitor transitions at Nanao Station

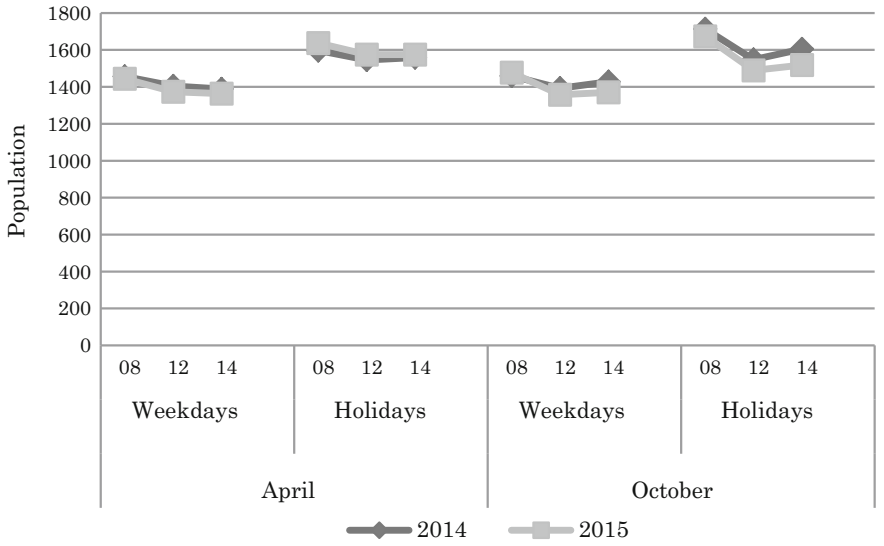


Fig. 9 Visitor transitions at Yamanaka Hot Springs

On the other hand, many elderly people (over 60 years old) visited the hot springs and the Noto Peninsula. I presume that the local people account for a large proportion of these visitors (Fig. 12).

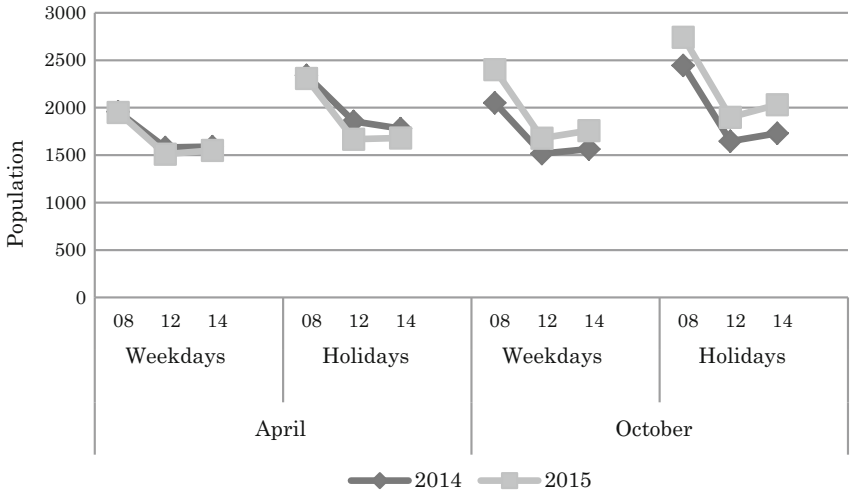
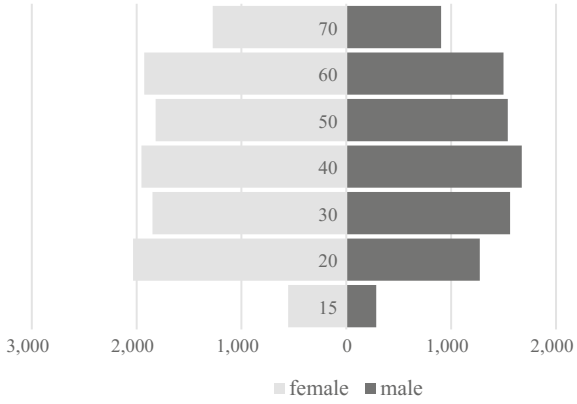


Fig. 10 Visitor transitions at Wakura Hot Springs

Fig. 11 Visitors' gender distribution at Kenrokuen (12:00 a.m. –1:00 p.m. on holidays in October 2015)



4.2.2 A Comparison of Kanazawa and Toyama

A comparison of visitors at Kanazawa Station with those at Toyama Station found that Kanazawa Station attracted visitors from a wider area of Japan (see Figs. 13 and 14). Mobile phone users around Kanazawa Station were from 235 municipalities, including Ishikawa Prefecture, whereas those at Toyama Station were from 43 municipalities, including Toyama Prefecture (12:00 a.m.–1:00 p.m. on holidays in October 2015).

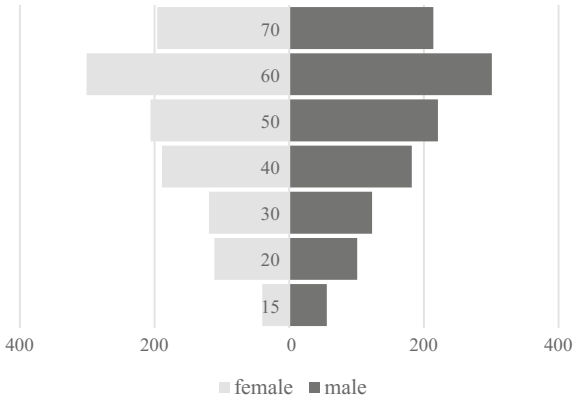


Fig. 12 Visitors' gender distribution at Wajima (12:00 a.m.–1:00 p.m. on holidays in October 2015)

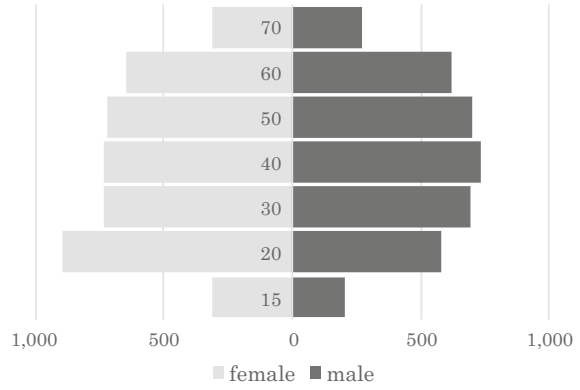


Fig. 13 Visitors' gender distribution at Kanazawa Station (12:00 a.m.–1:00 p.m. on holidays in October 2015)

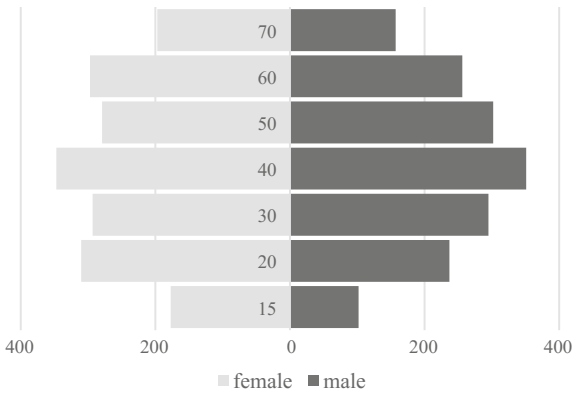


Fig. 14 Visitors' gender distribution at Toyama Station (12:00 a.m.–1:00 p.m. on holidays in October 2015)

Table 2 Visitors' residential distribution at Kanazawa Station and Toyama Station (12:00 a.m.–1:00 p.m. on holidays in October 2014 and 2015)

Kanazawa station			Toyama station		
Residence	2015	2014	Residence	2015	2014
Toyama, Toyama Pref.	131	147	Kanazawa, Ishikawa Pref.	67	60
Takaoka, Toyama Pref.	93	94	Fukui, Fukui Pref.	18	13
Fukui, Fukui Pref.	91	80	Setagaya-ku, Tokyo	16	n/a
Myoko, Niigata Pref.	47	n/a	Suginami-ku, Tokyo	15	n/a
Nanto, Toyama Pref.	34	31	Yamagata, Yamagata Pref.	14	n/a
Setagaya-ku, Tokyo	34	17	Takayama, Gifu Pref.	14	n/a
Imizu, Toyama Pref.	34	42	Nakamura-ku, Nagoya, Aichi Pref.	13	n/a
Oyabe, Toyama Pref.	33	39	Minami-ku, Niigata, Niigata Pref.	13	n/a
Sakai, Fukui Pref.	32	30	Gifu, Gifu Pref.	12	n/a
Omachi, Nagano Pref.	30	n/a	Hakusan, Ishikawa Pref.	12	20
Ota-ku, Tokyo	28	12	Ota-ku, Tokyo	12	n/a
Tsubame, Niigata Pref.	28	n/a	Himeji, Hyogo Pref.	12	n/a
Bunkyo-ku, Tokyo	27	n/a	Nagano, Nagano Pref.	12	n/a
Nagano, Nagano Pref.	24	n/a	Nerima-ku, Tokyo	11	n/a
Hiratsuka, Kanagawa Pref.	24	n/a	Shinagawa-ku, Tokyo	11	n/a
Sanda, Hyogo Pref.	23	n/a	Hida City, Gifu Pref.	11	10
Tonami, Toyama Pref.	23	26	Joetsu, Niigata Pref.	11	n/a
Suginami-ku, Tokyo	22	16	Kawaguchi, Saitama Pref.	10	n/a
Nerima-ku, Tokyo	22	10	Adachi-ku, Tokyo	10	n/a
Tsu, Mie Pref.	22	n/a	Takatsuki, Osaka	10	n/a

Note The figures above do not include visitors from Ishikawa Pref. for Kanazawa Station and Toyama Pref. for Toyama Station
Numbers less than ten are represented as "n/a"

Although the Hokuriku Shinkansen stops at both stations, the results suggest that Kanazawa has been more successful so far in attracting visitors (see Table 2 and Fig. 15). The number of visitors from Tokyo (gray column) increased in both the cities. Despite the fact that Toyama is nearer to Tokyo than Kanazawa, the latter successfully attracted more visitors from Tokyo.

Regarding the two prefectures, Ishikawa and Toyama, they both do have many wonderful tourist attractions. However, as for the two cities, it seems that Kanazawa is more attractive to tourists.

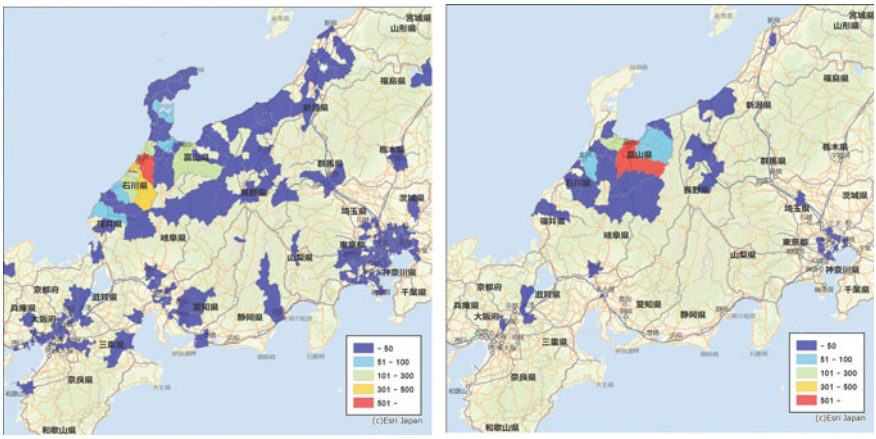


Fig. 15 Visitors' residential distribution at Kanazawa Station and Toyama Station (12:00 a.m.–1:00 p.m. on holidays in October 2015)

5 Toward Utilizing Big Data in Tourism

Tourism is likely to become increasingly important to local economies. However, tourism industry has been suffering significant volatility in demand depending on the season and day of the week. In addition, there was significant loss of business opportunities because of congestion during the busy season.

To cope with such volatility, various events have been held to eliminate the seasonal gap. Many events are newly launched. To date, it has been difficult to accurately grasp the extent to which these events attract visitors and the types of people who visit. However, by employing the recently provided Information and Communication Technology (ICT) services, it is possible to verify the number and characteristics of visitors to a particular event.

In this chapter, I attempted to identify the number of visitors at two points in time at various places in Japan and their characteristics using the location data of mobile phone users collected by the mobile phone company. As explained above, the opening of the Hokuriku Shinkansen increased the number of visitors to many areas. However, it also led to fewer visitors in some other areas. Its positive effect was remarkable in Kanazawa.

Numerous events have been recently held in Japan to attract visitors. In addition to using the *Mobile Kukan Toukei*, combining it with other ICT services, such as Google Trends, can help better predict the number of visitors at new events. Specifically, by combining the *Mobile Kukan Toukei* and the transition of the search results for a particular tourist destination, it would be possible to more accurately predict the number of tourists. If we can realize more accurate demand forecasting, it would be possible to optimize the necessary goods and number of non-regular employees

in advance. Moreover, understanding consumers' characteristics beforehand could enable us to optimize the services, which could influence customer satisfaction.

Note

"Mobile Kukan Toukei" is a trademark of NTT DOCOMO, Inc.
(*) NTT DOCOMO's "Mobile Kukan Toukei" services are only available to subscribers in Japan.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP15K01970.

References

1. Dolnicar, S., & Ring, A. (2014). Tourism marketing research: Past, present, and future. *Annals of Tourism Research, 47,* 31.
2. Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28,* 45–59.
3. Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations—A case from Sweden. *Journal of Destination Marketing & Management, 3*(4), 198–208.
4. Xiang, Z., Schwartz, Z., Gerdes, J. H., Jr., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44,* 120–129.
5. Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management, 29*(3), 469–485.
6. Japan Tourism Agency, Keitaidenwa kara erareru ichijouhou tou wo katuyousita hounichi gaikokujin doutaichousa houkokusho [Foreign visitors' dynamics research report utilizing mobile phone location information]. Retrieved from http://www.mlit.go.jp/common/001080545.pdf (2014).
7. Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications, 40*(8), 3299–3311.
8. Gao, H., & Liu, F. (2013). Estimating freeway traffic measures from mobile phone location data. *European Journal of Operational Research, 229*(1), 252–260.
9. Steenbruggen, J., Tranos, E., & Nijkamp, P. (2015). Data from mobile phone operators: a tool for smarter cities? *Telecommunications Policy, 39*(3–4), 335–346.
10. Okinawa Prefecture Culture, Sports, and Tourism Department, Senryakuteki repeater souzou jigyou houkokusho [Report on strategic creation of repeaters]. Retrieved from http://www.pref.okinawa.jp/site/bunka-sports/kankoseisaku/kikaku/report/houkokusixyo/documents/07dairokusiyou.pdf (2013).

C-LASSO Estimator for Generalized Additive Logistic Regression Based on B-Spline

Pakize Taylan and Gerhard Wilhelm Weber

1 Introduction

One of the most important statistical and data mining problems is that of regressing and classification a binary response variable Y that takes values representing success (1) and failure (0) or, more generally, the presence or absence of an attribute of interest. It is assumed that Y has a binomial distribution with binomial denominator n and probability p. This kind of a response variable can be encountered in real life and every field of science.

The nonparametric logistic regression model [19] is appropriate for such a statistical problem and it is the most widely used model for explicitly modeling the probability of an event and classifying binary response variable.

The logistic regression model for any vector $\boldsymbol{X} = \boldsymbol{x}$ given as

$$\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{f(\boldsymbol{x})}}{1 + e^{f(\boldsymbol{x})}} = p(\boldsymbol{x}). \tag{1.1}$$

Here, $f(\boldsymbol{x})$ can be estimated by using different estimation methods. One of the famous estimation methods is based on the linear regression [34], or a smoother based on the penalized log-likelihood criterion [8]. In the latter case, estimation of the conditional probability $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ can be obtained. A smoother may be used in estimating $f(\boldsymbol{x})$ then, the smooth estimation of $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ can be

P. Taylan (✉)
Science Faculty, Dicle University, 21280 Diyarbakır, Turkey
e-mail: ptaylan@dicle.edu.tr

G. W. Weber
Faculty of Management Engineering, Poznan Technology University, Poznan, Poland
e-mail: gerhard.weber@put.poznan.pl

G. W. Weber
IAM, METU, Ankara, Turkey

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_10

obtained and it is can be used for classification [6].

The score equations for estimation problem may be obtained from the penalized maximum likelihood estimation [11], given by the following optimization problem:

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \sum_{i=1}^{N} \log p(\boldsymbol{x}_i) - \alpha R(\boldsymbol{\theta}), \tag{1.2}$$

where $p(\boldsymbol{x}_i) = \Pr(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x}_i)$, $R(\boldsymbol{\theta})$ is a chosen regularization term, which forces the *complexity* of the model to be small, and α is a penalty parameter. Here, $(y_i, \boldsymbol{x}_i)$ represents data for ith data case $(i = 1, 2, \ldots, N)$. Now, the penalized log-likelihood criterion based on the binomial distribution to estimate $f(\boldsymbol{x})$ is given as follows:

$$l(f, \lambda) = \sum_{i=1}^{N} \left[y_i f(\boldsymbol{x}_i) - \log\left(1 + e^{f(\boldsymbol{x}_i)}\right) \right] - \tfrac{1}{2}\lambda \int_{Q} \left(D^2 f\right)^2 (\boldsymbol{t}) d\boldsymbol{t}. \tag{1.3}$$

Here, $\left(D^2 f\right)(\boldsymbol{t})$, is a cumulative term established on second-order partial derivative of the function $f(\boldsymbol{t})$ with respect to components of the vector $\boldsymbol{x}$, and the Q is domain of integration which encompasses all the input data $\boldsymbol{x}_i (i = 1, 2, \ldots, N)$. This integral is considered as a regularization term $R(\boldsymbol{\theta})$. Furthermore, $\alpha = \lambda/2$ is a constant regularization or smoothing parameter. The first term is a measure of proximity to the data, while the second one penalizes curvature of $f(\boldsymbol{t})$ according to the curvature severity of the function. Here, λ constructs a balance between the first and the second part.

In our study, we will estimate the probabilities $p(\boldsymbol{x}_i)$ based on a generalized additive model [15, 16] in which the model values $f(\boldsymbol{x}_i)$ are approximated in terms of sums of univalues B-spline $f_j(x_j)$. Then, we will construct LASSO type regularization [32] problem, but we try to solve it by Conic Quadratic Programming [4] and call it C-LASSO.

The second-order cone programming problem (SOCP) can be considered as an example of a quadratic programming. The standard form for a Second Order Cone Programming (SOCP) [4, 22] problem given as

$$\underset{\boldsymbol{x}}{\text{minimize}} \ \boldsymbol{c}^T \boldsymbol{x}, \quad \text{where} \quad \boldsymbol{A}_i \boldsymbol{x} - \boldsymbol{b}_i \in L^{n_i} \ (i = 1, 2, \ldots, k). \tag{1.4}$$

2 Generalized Additive Model (GAM)

Given p predictors or input $X_1, X_2, \ldots, X_p$, and a response Y for the input variable, then, a *generalized additive model* (*GAM*) [15, 16] has the main form

$$\eta(\boldsymbol{X}) = \mathrm{G}(\mu) = \theta_0 + \sum_{j=1}^{p} f_j(X_j), \tag{2.1}$$

where $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x})$ is conditional mean of Y, G is a 'link function' that linked $\mu = E(Y|\boldsymbol{X} = \boldsymbol{x})$ to the predictors $X_1, X_2, \ldots, X_p$, the functions f_j are smooth function of X_j and unspecified ("nonparametric") and θ_0 is a unknown parameter. In addition, GAM usually makes the distributional assumptions that Y has an exponential family density $\varphi_Y(y, \varsigma, \rho)$ with the natural parameter ς and the dispersion parameter ρ. GAMs are extensions of generalized linear models [23] in which the linear form $\sum \theta_j X_j$ is replace by a sum of smooth functions $\sum f_j(X_j)$. The incorporation of θ_0 as some average outcome allows us to assume $E\big(f_j(X_j)\big) = 0 (j = 1, 2, \ldots, p)$.

The estimation problem, using GAMs, needs that $\boldsymbol{\theta} = \big(\theta_0, f_1, \ldots, f_p\big)^T$ be considered as the unknown parameter vector to be estimated. Often, the unknown functions f_j can be obtained with the help of finite-dimensional space of functions. Generally, these functions are considered to be combination of spline basis functions approximating the data [9]. The spline orders can be determined appropriately according to the variation and density features of the corresponding data in the x and y components, respectively. Then, the problem of determining $\boldsymbol{\theta}$ also turn to a finite-dimensional parameter estimation problem. The common algorithm which is used to estimate GAMs consists of a combination of basic backfitting and basic local scoring algorithms; therefore, estimating GAMs is composed of two basic loops. The weighted backfitting algorithm (inner loop) is used to estimate functions f_j within each step of the local scoring algorithm (outer loop) and this process is continued up to convergence is attained. Then, new weights are computed by using estimates of functions f_j obtained from this algorithm, and the next iteration for the local scoring algorithm begins. Given a vector for coefficient, $\boldsymbol{\theta}^0$, a vector of linear predictor $\boldsymbol{\eta}^0 = \big(\eta_1^0, \ldots, \eta_N^0\big)^T$ and $\boldsymbol{\mu}^0 = \big(\mu_1^0, \ldots, \mu_N^0\big)^T$, the framework of the *local scoring algorithm* procedure looks as follows [15]:

***Step* 1**. $\theta_0 = G\Big(\sum_{i=1}^N y_i/n\Big)$; $f_j^0 = 0\,(j = 1, \ldots, p)$, $m = 0$.

***Step* 2**. Iterate: $m \leftarrow m + 1$.

Set the adjusted dependent variable as

$$s_i = \eta_i^0 + \big(y_i - \mu_i^0\big)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0 \text{ with } \eta_i^0 = \theta_0^0 + \sum_{j=1}^{p} f_j^0\big(x_{ij}\big),\ \mu_i^0 = G^{-1}\big(\eta_i^0\big).$$

Set the weights:

$$w_i^{-1} = (\partial \eta_i/\partial \mu_i)_0^2\big(C_i^0\big),$$

where C_i is the variance matrix for Y_i. Additive model is fitted to s_i, to get estimation of f_j^1, η_i^1 and μ_i^1, respectively.

Then, calculate the convergence depend on two neighbouring iterations

$$\Delta\left(\eta^1, \eta^0\right) = \left(\sum_{j=1}^{p} \left\| f_j^1 - f_j^0 \right\|_2\right) \Big/ \left(\sum_{j=1}^{p} \left\| f_j^0 \right\|_2\right)$$

***Step* 3**. Reiterate step 2 changing η^0 by η^1 till $\Delta\left(\eta^1, \eta^0\right)$ is less than some small threshold.

Here, $\|\boldsymbol{f}\|_2 := \left\| \left(f(x_{ij}), \ldots, f(x_{Nj})\right)^T \right\|_2$ is the length of $\boldsymbol{f}$ that evaluated at the N sample points [16].

GAM is a powerful tool in regression and classification setting. It may be use for modeling of class probabilities for logistic regression. The g*eneralized additive logistic regression model* (*GALRM*) [15] has the form

$$\log \frac{\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{\Pr(Y = 0|\boldsymbol{X} = \boldsymbol{x})} = f(\boldsymbol{X}) = \theta_0 + \sum_{j=1}^{p} f_j\left(X_j\right), \tag{2.2}$$

where $\boldsymbol{X} = \left(X_1, X_2, \ldots, X_p\right)^T$ is considered to consist of prognostic factors and f_j $(j = 1, 2, \ldots, p)$ are unspecified smooth functions. Our aim is to model the term $\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ by using prognostic factors. For this reason, the parameter θ_0 and the smooth functions f_j $(j = 1, 2, \ldots, p)$ should be estimated. Generally, these functions are assessed by the local scoring algorithm [15] which is a backfitting algorithm within a Newton–Raphson procedure. Given a current N observation values y_i that can be coded as 0 or 1, then the cycle of backfitting algorithm is given as follows:

***Step* 1.** Calculate initial values: $\hat{\theta}_0 = \log\left(\frac{\bar{y}}{1-\bar{y}}\right)$, where $\bar{y} = \text{ave}\{y_i | i = 1, 2, \ldots, N\}$ is the sample proportion of ones and $\hat{f}_j \equiv 0$ $(j = 1, 2, \ldots, p)$.

***Step* 2**. Describe $\hat{\eta}_i \equiv \hat{\theta}_0 + \sum_{j=1}^{p} f_j(x_{ij})$ $(j = 1, \ldots, p)$ and $\hat{p}_i = 1/\left[1 + \exp\left(-\hat{\eta}_i\right)\right]$.

Iterate:

(i) Establish a study target variable: $v_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}$.

(ii) Establish weights: $w_i = \hat{p}_i\left(1 - \hat{p}_i\right)$.

(iii) Apply an additive model to the targets v_i with w_i, using a weighted backfitting algorithm and obtain new estimates for $\hat{\theta}_0$ and $\hat{f}_j$ $(j = 1, \ldots, p)$.

***Step* 3**. Reiterate step 2 till the change in the functions is less than a predetermined some small threshold.

3 LASSO Estimate

3.1 LASSO Problem for GALRM with B-Spline

In Sect. 2, we mentioned the local scoring algorithm to obtain unspecified functions $\hat{f}_j$ and $\hat{\eta}_i$. In this section, we propose an alternative method to local scoring algorithm by B-spline to estimate p_i probabilities and functions f_j.

Now we consider

$$\Pr(Y_i = 1|\boldsymbol{X} = \boldsymbol{x}_i) = \frac{e^{f(\boldsymbol{x}_i)}}{1 + e^{f(\boldsymbol{x}_i)}}, \tag{3.1}$$

where

$$f(\boldsymbol{x}_i) = \theta_0 + \sum_{j=1}^{p} f_j(x_{ij}), \tag{3.2}$$

for $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T$. If we have data consisting of N realizations of a binary outcome variable Y at p prognostic factors, then the penalized log-likelihood criterion [11] based on the Binomial distribution to estimate $f(\boldsymbol{x}_i)$ is given as

$$\begin{aligned} l(f, \boldsymbol{\lambda}) &= \sum_{i=1}^{N} \left[y_{\mathrm{i}} \log p(\boldsymbol{x}_{\mathrm{i}}) + (1 - y_{\mathrm{i}}) \log(1 - p(\boldsymbol{x}_{\mathrm{i}}))\right] - \tfrac{1}{2} \sum_{j=1}^{p} \lambda_j \int_{Q_j} \left(f_j''(t_j)\right)^2 dt_j \\ &= \sum_{i=1}^{N} \left[y_{\mathrm{i}} f(\boldsymbol{x}_{\mathrm{i}}) - \log(1 + e^{f(\boldsymbol{x}_i)})\right] - \tfrac{1}{2} \sum_{j=1}^{p} \lambda_j \int_{Q_j} \left(f_j''(t_j)\right)^2 dt_j. \end{aligned} \tag{3.3}$$

Here, we consider a regularization term based on the sum of squared second-order partial derivative of functions f_j, which is a special case for $\left(D^2 f\right)(\boldsymbol{t})$. Furthermore, $Q = \prod_{j=1}^{p} Q_j$ with intervals Q_j. If Eq. (3.2) is used in the definition of $l(f, \lambda)$, then $l(f, \lambda)$ may be written as follows:

$$\begin{aligned} l(f, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \left[y_{\mathrm{i}} \left(\theta_0 + \sum_{j=1}^{p} f_j\left(x_{ij}\right) \right) - \log\left(1 + e^{\theta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{x}_{ij})} \right) \right] \\ - \tfrac{1}{2} \sum\nolimits_{j=1}^{p} \lambda_j \int_{Q} \left(f_j''\left(t_j\right)\right)^2 dt_j. \end{aligned} \tag{3.4}$$

A maximizer for $l(f, \lambda)$ can be obtained by addressing different kinds of spline bases functions, including the truncated power basis which is very simple. But it is not too attractive numerically since calculations of spline by means of the truncated

polynomials may reveal instability problem, the design matrix can consist of large values, and there may be collinearity between the columns of the design matrix, also powers of large numbers can cause to sharp rounding errors. The *B-spline* basis provides efficient computations even when the number of knots K is large. Therefore, we propose to use an additive combination of k-order B-spline with knots at the unique values x_{ij} instead of generic or unspecificied functions f_j.

Zero- and k-degree B-spline are defined by

$$B_{j,0}(z) = \begin{cases} 1, \text{ if } \ z_j \le z < z_{j+1}, \\ 0, \text{ otherwise}, \end{cases} \tag{3.5}$$

and

$$B_{j,k}(z) = \frac{z - z_j}{z_{j+k} - z_j} B_{j,k-1}(z) + \frac{z_{j+k+1} - z}{z_{j+k+1} - z_{j+1}} B_{j+1,k-1}(z) \ (k \ge 1, \ k \in \mathbb{N}); \tag{3.6}$$

for $k \ge 2$ its derivative [10, 26], wherever defined, is given by

$$\frac{d}{dt} B_{j,k}(z) = \frac{k}{z_{j+k} - z_j} B_{j,k-1}(z) - \frac{k}{z_{j+k+1} - z_{j+1}} B_{j+1,k-1}(z). \tag{3.7}$$

B-spline basis functions overlap with neighboring ones. The amount of this overlap increases as the degree of basis functions increases.

Let us abbreviate $B_{s,k}(z) := B_s(z)$ for the simplicity of our formula and write the functions $f_j(z)$ as

$$f_j(z) = \sum_{s=1}^{r} \theta_s^j B_s(z), \ \text{ if } x_{i,j-1} \le z \le x_{i,j+1},$$

in Eq. (3.3). Then, the penalized log-likelihood function $l(f, \lambda)$ will take the following form:

$$\begin{aligned} l(f, \boldsymbol{\lambda}) = & \sum_{i=1}^{N} \left[y_i \left(\theta_0 + \sum_{j=1}^{p} \sum_{s=1}^{r} \theta_s^j B_s(x_{ij}) \right) - \log \left(1 + e^{\theta_0 + \sum_{j=1}^{p} \sum_{s=1}^{r} \theta_s^j B_s(x_{ij})} \right) \right] \\ & - \tfrac{1}{2} \sum_{j=1}^{p} \lambda_j \int_{a_j}^{b_j} \left[\sum_{s=1}^{r} \theta_s^j B_s''(z_j) \right]^2 dz_j \\ = & \sum_{i=1}^{N} \left[y_i \boldsymbol{H_i \theta} - \log \left(1 + e^{\boldsymbol{H_i \theta}} \right) \right] - \tfrac{1}{2} \sum_{j=1}^{p} \lambda_j \boldsymbol{\theta}^{jT} \boldsymbol{\Omega}_j \boldsymbol{\theta}^j. \end{aligned} \tag{3.8}$$

Here, $\boldsymbol{\theta} = \left(\theta_0, \boldsymbol{\theta}^{1\mathrm{T}}, \ldots, \boldsymbol{\theta}^{p^T} \right)^{\mathrm{T}}$ is the aligned $(rp + 1)$-vector of unknown parameters whose lth sub-block is $\boldsymbol{\theta}^j = \left(\theta_1^j, \theta_2^j, \ldots, \theta_r^j \right)^T$, $\boldsymbol{H}_i$ is a row $(1 + rp)$-vector constructed by $\boldsymbol{H}_i = (1 \, \boldsymbol{B}_i)$, where $\boldsymbol{B}_i = \left(\boldsymbol{B}_i^1, \ldots, \boldsymbol{B}_i^p \right)$ is the aligned sub-vector with

$\boldsymbol{B}_i^j = \left(B_1(x_{ij}), B_2(x_{ij}), \ldots, B_r(x_{ij})\right)$, $\boldsymbol{\Omega}_j$ is the symmetric $(r \times r)$-matrix whose (s, k)th element is $\left(\boldsymbol{\Omega}_j\right)_{sk} = \int_{a_j}^{b_j} B_s''(t_j) B_k''(t_j) dt_j$ and the interval bounds of $Q_j = [a_j, b_j]$ depend on the coordinate j.

To make the calculations of

$$\int_{a_j}^{b_j} \left[\sum_{s=1}^{r} \theta_s^j B_s''(z_j)\right]^2 dz_j$$

more easy, the following approximation can be used:

$$\left(\boldsymbol{\Omega}_j\right)_{sk} = \int_{a_j}^{b_j} B_s''(z_j) B_k''(z_j) dz_j \cong \sum_{i=1}^{N-1} B_k''(x_{ij}) B_k''(x_{ij}) \Delta x_{ij}, \tag{3.9}$$

where $\Delta x_{ij} := \left(x_{i+1j} - x_{ij}\right)$ $(i = 1, 2, \ldots, N-1)$.

In our problem, there is a finite sequence of *penalty* parameters $\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)^T$. That situation makes it very difficult to solve the problem. For this reason, let us make a uniform penalization by taking into consideration the same λ for each derivative term. Then, approximation of $l(f, \boldsymbol{\lambda})$, can be rearranged as

$$l(f, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \left[y_i \boldsymbol{H}_i \boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{H}_i \boldsymbol{\theta}}\right)\right] - \tfrac{1}{2} \lambda \boldsymbol{\theta}^T \boldsymbol{\Omega} \boldsymbol{\theta}, \tag{3.10}$$

where $\boldsymbol{\Omega}$ is a block matrix whose first row and first column consist of zeros, and the diagonal elements consists of the matrix $\boldsymbol{\Omega}_j$ except the first row and first column. Taking the first-order derivative of $l(f, \lambda)$ with respect to $\boldsymbol{\theta}$, we obtain the gradient of l:

$$\nabla l(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{H}^T (\boldsymbol{y} - \boldsymbol{p}) - \lambda \boldsymbol{\Omega} \boldsymbol{\theta}, \tag{3.11}$$

where $\boldsymbol{p}$ is an N-vector whose ith element is $p(\boldsymbol{x}_i)$, and $\boldsymbol{H}$ is an $N \times (rp + 1)$-matrix whose ith row consist of the $\boldsymbol{H}_i$ vector. The score equations

$$\nabla l(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{0},$$

can be solved by the Newton–Raphson algorithm [21] which can be obtained from a first-order linear approximation to the gradient $\nabla l(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ and requires the second-order derivatives or the Hessian matrix. Taking the second-order partial derivative of $l(f, \lambda)$ with respect to $\boldsymbol{\theta}$, the Hessian matrix for score equations can be obtained as

$$\nabla^2 l(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = -\boldsymbol{H}^T \boldsymbol{W} \boldsymbol{H} - \lambda \boldsymbol{\Omega}, \tag{3.12}$$

where $\boldsymbol{W}$ is a diagonal matrix of weights $p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i))$. The score equations are similar to those from univariate logistic regression. Then, as a result of score Eq. (3.12), the solution $\hat{\boldsymbol{\theta}}$ is obtained:

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{H}^T \boldsymbol{W} \boldsymbol{H} + \lambda \boldsymbol{\Omega}\right)^{-1} \boldsymbol{H}^T \boldsymbol{W} \boldsymbol{a}, \tag{3.13}$$

where $\boldsymbol{a} = \left(\boldsymbol{H}\boldsymbol{\theta}^m + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p})\right)$ is adjusted response. As it can be seen that the solution $\hat{\boldsymbol{\theta}}$ is a generalized ridge regression [33]. Let us consider the vector

$$\boldsymbol{f}(\boldsymbol{x}) = \left[f(\boldsymbol{x}_1), f(\boldsymbol{x}_2) \ldots f(\boldsymbol{x}_2)\right]^T = \boldsymbol{H}\boldsymbol{\theta}.$$

Then, the fitted value $\hat{\boldsymbol{f}}$ may be written as

$$\hat{\boldsymbol{f}} = \boldsymbol{S}_{\lambda \mathrm{W}} \boldsymbol{a}, \tag{3.14}$$

where

$$\boldsymbol{S}_{\lambda,\mathrm{W}} = \boldsymbol{H}\left(\boldsymbol{H}^T \boldsymbol{W} \boldsymbol{H} + \lambda \boldsymbol{\Omega}\right)^{-1} \boldsymbol{H}^T \boldsymbol{W}, \tag{3.15}$$

is a weighted smoothing spline to the working response z and it is a positive semidefinite $N \times N$-matrix. Here, the penalty parameter λ traces out a set of ridge solutions and it should be choosen so that for which the coefficients are not rapidly changing and have 'sensible' signs. Details and discussions about the penalty parameter can be found in the paper of Hoerl and Kennard [18]. Also, this parameter can be obtained by automatic methods using techniques such as generalized cross-validation [13]. This technique has the advantage of giving an appropriate approach to leave-one-out cross-validation, for linear regression by using squared-error loss. Let us go back to Eq. (3.13) to get the optimal value of the penalty parameter by using generalized cross-validation. It can be seen that the estimation of $\boldsymbol{\theta}$ is the solution of penalized residual sum of squares (PRSS) optimization problem with rp basis functions, given as

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{rp+1}}{\text{minimize}} \quad (\boldsymbol{u} - \boldsymbol{H}\boldsymbol{\theta})^T \boldsymbol{W} (\boldsymbol{u} - \boldsymbol{H}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\theta}$$

or, equivalently,

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{rp+1}}{\text{minimize}} \quad \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{S}_1 \boldsymbol{\theta}\|_2^2, \tag{3.16}$$

where $\boldsymbol{W} = \boldsymbol{L}^T \boldsymbol{L}$ and $\boldsymbol{\Omega} = \boldsymbol{S}_1^T \boldsymbol{S}_1$ are Cholesky decompositions [14] of $\boldsymbol{W}$ and $\boldsymbol{\Omega}$, respectively, $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{H}$, $\boldsymbol{b} = \boldsymbol{L}\boldsymbol{a}$ and $\|\cdot\|_2$ stands for Eucludian norm. The penalty

function $\|\boldsymbol{S}_1\boldsymbol{\theta}\|_2^2$ used in Eq. (3.16) is only one of the available options. Different types of penalty functions with positive and negative characteristics can be written, like the method of *Least Absolute Shrinkage and Selection Operator*, in short: *LASSO* [32]. LASSO method has a superior property to Tikhonov regularization or ridge regression by that it uses a L_1-penalty. Ridge regression is a continuous method of shrinking the coefficients $\boldsymbol{\theta}$ and is more consistent. However, it does not state coefficients to 0 and, hence, does not obtain an easily interpretable model. The main motivation is that LASSO typically yields a sparse vector and shrinks some coefficients and states others to 0; hence, it tries to hold good properties of both subset selection and ridge regression. Therefore we consider to use, the L_1-penalty in the maximum likelihood estimation to obtain a better interpretable solution than by a logistic regression. The L_1-regularized logistic regression problem based on the generalized additive model and B-spline can be written as

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^{rp}}{\text{minimize}} \quad h(\boldsymbol{\theta}) &:= \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \\ &= L(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1, \end{aligned} \tag{3.17}$$

where $L(\boldsymbol{\theta}) := \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2$ and $\|\boldsymbol{\theta}\|_1 := \sum_{v=1}^{rp} |\theta_v|$. It is very convenient to not penalize the intercept term, and standardize the predictors for the penalty to be meaningful when a regularization technique is used. Therefore, we do not penalize the intercept term in the problem of Eq. (3.17). Since the objective function is nondifferentiable when a component of $\boldsymbol{\theta}$ contains values of 0, this does not permit the use of standard unconstraint optimization methods.

3.2 Solution Methods for LASSO Problem

There are several methods to solve the problem in Eq. (3.17) like Quadratic Programming, iterated ridge regression [2] and methods which use subgradient strategies [30]. It needs 2^{rp} constraint functions to solve problem in Eq. (3.17) by Quadratic Programming; therefore, one cannot iterate over all the constraints generated by this expansion for non-trivial values of rp.

As mentioned above, we have a nondifferentiable objective function in Eq. (3.17). Such a nondifferentiable objective is considered as a nonsmooth optimization problem [3] treated by using subgradients of the objective function at nondifferentiable points. In this type of optimization problem, the fact that the $\mathbf{0}$-vector has to belong to the subdifferential $\partial h(\boldsymbol{\theta})$ defined as the set containing all subgradients at some $\boldsymbol{\theta}$ [12], is a necessary condition for a parameter vector $\boldsymbol{\theta}$ to be local minima. The subdifferential of the absolute value function $|\theta_v|$ is given by the signum function $\text{sign}(\theta_v)$ defined as

$$\text{sign}(\theta_v) = \begin{cases} \text{sign of } \theta_v, & \theta_v \neq 0, \\ \gamma \in [-1, 1], & \theta_v = 0. \end{cases}$$

For our problem, at a local minimize $\tilde{\boldsymbol{\theta}}$ of the objective function in Eq. (3.17), the following first-order optimality conditions [28] are considered; we focus on the case of a single active constraint $\theta_v = 0$:

$$\begin{cases} \nabla_v L(\tilde{\boldsymbol{\theta}}) + \lambda \text{sign}(\tilde{\theta}_v) = 0, & \left|\tilde{\theta}_v\right| > 0, \\ \left|\nabla_v L(\tilde{\boldsymbol{\theta}})\right| \leq \lambda, & \tilde{\theta}_v = 0. \end{cases} \tag{3.18}$$

These conditions can be used to define the following subgradient $\nabla_v^s h(\boldsymbol{\theta})$ for each θ_v whose negation represents the coordinate-wise direction of maximum descent:

$$\nabla_v^s h(\boldsymbol{\theta}) := \begin{cases} \nabla_v L(\boldsymbol{\theta}) + \lambda \text{sign}(\theta_v), & |\theta_v| > 0, \\ \nabla_v L(\boldsymbol{\theta}) + \lambda, & \theta_v = 0, \nabla_v L(\boldsymbol{\theta}) < -\lambda, \\ \nabla_v L(\boldsymbol{\theta}) - \lambda, & \theta_v = 0, \nabla_v L(\boldsymbol{\theta}) > \lambda, \\ 0, & \theta_v = 0, \lambda \leq \nabla_v L(\boldsymbol{\theta}) \leq -\lambda. \end{cases} \tag{3.19}$$

For a suboptimal $\boldsymbol{\theta}$, this subgradient will yield a descent direction on the objective function $h(\boldsymbol{\theta})$, and these methods optimize over one variable at a time. In those methods, it is assumed that $\nabla^2 h(\boldsymbol{\theta}) := \nabla^2 L(\boldsymbol{\theta})$, even though this is not strictly true if any θ_v is 0. The nonsmooth optimization problem of Eq. (3.17) for logistic regression can be implemented by the Gauss-Seidel algorithm of Shevade and Keerthi [29].

The methods using subgradient are difficult for large-scale problems, as they may require so many coordinate updates that they become impractical. Therefore, we will consider methods that update multiple variables on each iteration. The *Iterated Ridge Regression* (IRR) [2] method is one of these methods that update multiple variables at each iteration and it is a differentiable approximation method to L_1-regularization. For our problem we will prefer the IRR and we will solve it by CQP. Iterated Ridge Regression method is based on the following approximation:

$$|\theta_v| \approx \frac{\theta_v^2}{\left|\theta_v^k\right|}, \tag{3.20}$$

where θ_v^k is the value from the previous iteration k. Substituting this approximation into the unconstrained formulation in Eq. (3.17), we can obtain an expression similar to least-squares estimation with a L_2-penalty (Ridge regression) as follows:

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{rp}}{\text{minimize}} \quad h(\boldsymbol{\theta}) := \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2 + \lambda \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{\theta}, \tag{3.21}$$

where $\boldsymbol{R}$ is diagonal matrix whose vth diagonal element is $\left|\theta_v^{\mathrm{k}}\right|^{-1}$, i.e.,

$\boldsymbol{R} = \text{diag}\left(\left|\theta_1^k\right|^{-1}, \left|\theta_2^k\right|^{-1}, \ldots, \left|\theta_{rp}^k\right|^{-1}\right) := \text{diag}\left(\left|\boldsymbol{\theta}^k\right|^{-1}\right)$. The solution of the problem of Eq. (3.21) for the kth step is

$$\boldsymbol{\theta}^{k+1} = \left(\boldsymbol{M}^T\boldsymbol{M} + \lambda \hat{\text{diag}}\left|\boldsymbol{\theta}^{\text{k}}\right|^{-1}\right)^{-1}\boldsymbol{M}^{\text{T}}\boldsymbol{b}. \tag{3.22}$$

However, we should note that this approximation will be numerically unstable when one of the θ_v^k approaches 0. To get rid of this problem, we may consider a generalized inverse [1, 25] of $\left|\boldsymbol{\theta}^k\right|$. Hence, the values too close to zero are removed from the estimation problem. However, a new problem arises with this inverse; zero-valued variables cannot move away from 0 and, thus, it could potentially lead to sub-optimal results if the initialization is inappropriate.

Let us use the Cholesky decomposition of $\boldsymbol{R} = \boldsymbol{S}_2^T\boldsymbol{S}_2$ and substitute it into Eq. (3.21); thus we get Eq. (3.21) as

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^{rp}}{\text{minimize}} \quad h(\boldsymbol{\theta}) := \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{S}_2\boldsymbol{\theta}\|_2^2. \tag{3.23}$$

Equation (3.23) is a special case of Ridge regression and we may find C-LASSO estimate by solving this differentiable optimization problem by CQP to be discussed in detail in the following section.

4 On Conic Optimization and Its Application in LASSO Problem for GALRM with B-Spline

4.1 Convex and Conic Optimization

Convex optimization deal with problems related with minimizing a convex function using a convex set. These problems, which have so many significant properties, like strong duality theory and the fact that any local minimum is a global minimum, often occur in various application fields. Such optimization problems provide advantages such as computationally easy working and theoretically efficient solution methods. Convex optimization contains various significant classes of problems such as semidefinite programming, second order cone programming, and geometric programming. Some information about convex optimization is given in the following by benefiting from [4, 31].

General form of a convex optimization problem defined as

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x}, \text{ for } \boldsymbol{x} \in U;$$

where, $\boldsymbol{c} \in \mathbb{R}^n$ and $U \subseteq \mathbb{R}^n$ is a convex set. *Linear programming* (*LP*), in which the objective and all constraint functions f_i $(i = 0, 1, \ldots, m)$ are linear, is the simplest case of a convex program:

$$\underset{\boldsymbol{u}\in\mathbb{R}^n}{\text{minimize}} \quad f_0(\boldsymbol{u}), \quad \text{where} \quad f_i(\boldsymbol{u}) \leq 0 \ (i = 1, 2, \ldots, m). \tag{4.1}$$

Such a problem can be written in the canonical form

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x}, \quad \text{where} \quad \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \in K := \mathbb{R}^n_+. \tag{4.2}$$

If, however, the objective or constraints are nonlinear, then we must take into account the nonlinearity in the corresponding function f_i in Eq. (4.1). It is easily seen [4] that a convex program (4.1) can be represented in the conic form similar to Eq. (4.2):

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x}, \quad \text{where } \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \in K. \tag{4.3}$$

here, $K \subseteq \mathbb{R}^N$ is a cone (closed, pointed, convex and with a nonempty interior), and $\mathbb{R}^n \to \mathbb{R}^N$, defined by $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ is a linear embedding.

Generally, convex programs depend on three generic cones $\boldsymbol{K}$ (in the second case referring to the Euclidean or ℓ_2 norm):

$$\textit{Nonnegative orthant}: \ \mathbb{R}^n_+ = \left\{\boldsymbol{x} \in \mathbb{R}^n \,|\, \boldsymbol{x} \geq 0\right\},$$

$$\textit{Direct products of Lorentz cone}: \ L^n = \left\{(\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R} \,|\, \|x\|_2 \ \leq \ t\right\},$$

$$\textit{Semidefinite cone}: \ S^n_+ = \left\{X \ \in S^n \,|\, X \succeq 0\right\},$$

they will give closer introduced below. The optimization problems based on these three cones can be solved by primal-dual interior point methods. These methods are very effective methods for *linear*, *conic quadratic* and *semidefinite* programming—all are examples of conic problems.

In the following sections, we shall pay attention to the class of *conic quadratic* problems which we apply to GALRM. For the cone underlying these problems, it can be describing explicitly the dual cone. Because in many cases, "duality" is very important for understanding of original models and converting it into equivalent forms better suited for numerical processing, etc.

4.2 Conic Quadratic Programming

The n-dimensional *ice-cream* (:= *second-order*, or *Lorentz*) cone L^n is defined by:

$$L^n = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n \,|\, x_n \geq \sqrt{x_1^2 + \cdots + x_{n-1}^2} \right\} \quad (n \geq 2).$$

A *conic quadratic problem* is a conic problem,

$$\underset{\boldsymbol{x}}{\text{minimize}} \;\; \boldsymbol{c}^T \boldsymbol{x}, \quad \text{where} \quad \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \in K, \tag{4.4}$$

for which the cone K is a direct product of several "*ice-cream cones*":

$$\begin{aligned} K &= L^{n_1} \times L^{n_2} \ldots \times L^{n_k} \\ &= \left\{ \left(\boldsymbol{s}[1]^T, \ldots, \boldsymbol{s}[k]^T\right)^T \;|\boldsymbol{s}[i] \in L^{n_i} \;\; (i = 1,2,\ldots,\; k) \right\}. \end{aligned} \tag{4.5}$$

From Eq. (4.5) we can see that a conic quadratic program is an optimization problem with a linear objective function and finitely many '*ice-cream constraints*'

$$\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{b}_i \in L^{n_i} \; (i = 1, 2, \ldots, \; k),$$

where

$$[\boldsymbol{A},\boldsymbol{b}] = \left[[\boldsymbol{A}_1;\boldsymbol{b}_1]^T, \ldots, [\boldsymbol{A}_k;\boldsymbol{b}_k]^T \right]^T$$

is the partition of the data matrix $[\boldsymbol{A},\; \boldsymbol{b}]$ corresponding to the partition of $\boldsymbol{s}$ in Eq. (4.5). Thus, our conic quadratic program can be written as

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{c}^T\boldsymbol{x}, \;\; \text{where } \boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{b}_i \in L^{n_i} \; (i = 1, 2, \ldots, \; k). \tag{4.6}$$

Sometimes, the relation $\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{b}_i \in L^{n_i}$ is also written in the form of a vector inequality, namely, $\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{b}_i \succeq_{L^{n_i}} 0$ or $\boldsymbol{A}_i\boldsymbol{x} \succeq_{L^{n_i}} \boldsymbol{b}_i$. This means a partial ordering. More generally, this kind of notation and partial order can be used in any a finite dimensional Euclidean space E where a good vector inequality '$\succeq$' is completely identified by the set K of '$\succeq$'-nonnegative vectors: $K = \{\boldsymbol{a} \in E \,|\, \boldsymbol{a} \succeq 0\}$, where $\boldsymbol{a} \succeq \boldsymbol{b} \Leftrightarrow \boldsymbol{a} - \succeq 0 (\Leftrightarrow \boldsymbol{a} - \boldsymbol{b} \in K)$. But the set K cannot be arbitrary. It must be a pointed convex cone. We note that every *pointed* convex cone K in E induces a partial ordering on E, given by '$\succeq_K$', where $\boldsymbol{a} \succeq_K \boldsymbol{b} \Leftrightarrow \boldsymbol{a} - \boldsymbol{b} \succeq_K 0 \Leftrightarrow \boldsymbol{a} - \boldsymbol{b} \in K$ [4].

Partitioning the data matrix $[\boldsymbol{A}_i;\boldsymbol{b}_i]$ by

$$[\boldsymbol{A}_i;\boldsymbol{b}_i] = \begin{bmatrix} \boldsymbol{D}_i & \boldsymbol{d}_i \\ \boldsymbol{p}_i^T & q_i \end{bmatrix},$$

with $\boldsymbol{D}_i$ being of the type $(n_i - 1) \times (\dim \boldsymbol{x})$, the problem can be written as

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \boldsymbol{c}^T\boldsymbol{x}, \quad \text{where} \quad \|\boldsymbol{D}_i\boldsymbol{x} - \boldsymbol{d}_i\|_2 \leq \boldsymbol{p}_i^T\boldsymbol{x} - q_i (i = 1, 2, \ldots, k). \tag{4.7}$$

Here, $\|\cdot\|_2$ is the Euclidean norm. This is a most explicit form of the conic problem and the one which we will use. In this form, $\boldsymbol{D}_i$ are matrices of the same row dimension as $\boldsymbol{x}$. Furthermore, the lengths of the column vectors $\boldsymbol{d}_i$ are the column dimensions of the matrices $\boldsymbol{D}_i$, and $\boldsymbol{p}_i$ are column vectors of the same dimension as $\boldsymbol{x}$; finally, q_i are reals. It can immediately be seen that (4.5) is indeed a cone, in fact a self-dual one: $K_* = K$[4].

As a result, the problem dual to Eq. (4.4) is

$$\underset{\boldsymbol{\lambda}}{\text{maximize}}\ \boldsymbol{b}^T\boldsymbol{\varpi}, \ \text{where}\ \boldsymbol{A}^T\boldsymbol{\varpi} = \boldsymbol{c}, \boldsymbol{\varpi} \in K \tag{4.8}$$

If we write $\boldsymbol{\varpi}$ as $\boldsymbol{\varpi} := \left(\boldsymbol{\varpi}_1^T, \boldsymbol{\varpi}_2^T, \ldots, \boldsymbol{\varpi}_k^T\right)^T$ with m_i-dimensional blocks $\boldsymbol{\varpi}_i$, then, the dual problem can be expressed as follows:

$$\underset{\varpi_1,\ldots,\varpi_n}{\text{maximize}} \sum_{i=1}^{k} b_i^T\boldsymbol{\varpi}_i, \quad \text{where} \sum_{i=1}^{k} \boldsymbol{A}_i^T\boldsymbol{\varpi}_i = \boldsymbol{c} \text{ and } \boldsymbol{\varpi}_i \in L^{n_i} (i = 1, 2, \ldots, k). \tag{4.9}$$

If it is considered $\boldsymbol{\varpi}_i = \left(\boldsymbol{\kappa}_i^T, \nu_i\right)^T$ with a scalar component ν_i and using the meaning of '$\geq_{L^n}$ 0', it can be shown that following form is the problem dual to Eq. (4.7):

$$\underset{(\boldsymbol{\kappa}_i),(\nu_i)}{\text{maximize}} \sum_{i=1}^{k} \left[\boldsymbol{\kappa}_i^T\boldsymbol{d}_i + \nu_i q_i\right], \ \text{where} \sum_{i=1}^{k} \left[\boldsymbol{D}_i^T\boldsymbol{\kappa}_i + \nu_i\boldsymbol{p}_i\right] = \boldsymbol{c}, \ \|\boldsymbol{\kappa}_i\|_2 \leq \nu_i \ (i = 1, 2, \ldots, k). \tag{4.10}$$

The design variables in Eq. (4.10) are column vectors $\boldsymbol{\kappa}_i$, having the same dimensions as the vectors $\boldsymbol{d}_i$, and reals ν_i $(i = 1, 2, \ldots, k)$. The programs Eqs. (4.7) and (4.10) are standard forms of a conic quadratic problem and of its dual.

4.3 *Application of Conic Quadratic Programming to LASSO Problem for GALRM with B-Spline*

Let us show how optimization over cones can be applied for a problem class from data mining and statistical learning which is motivated by real-world applications in, e.g., the financial sector, computational biology, health Sciences. For this reason, we formulate the optimization problem Eq. (3.23) as follows,

$$\begin{aligned} &\underset{\boldsymbol{\theta}\in\mathbb{R}^{rp}}{\text{minimize}} \quad G(\boldsymbol{\theta}) \\ &\text{subject to} \quad g(\boldsymbol{\theta}) \leq 0. \end{aligned} \tag{4.15}$$

Here, we have the objective function $G(\boldsymbol{\theta}) := \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2$ of least-squares and the constraint functions $g(\boldsymbol{\theta}) := \|\boldsymbol{S}_2\boldsymbol{\theta}\|_2^2 - C_2$. The original objective function can be moved to the list of constraints since it is not linear, and the equivalent problem is written in the following form,

$$\begin{aligned} &\underset{t_2,\boldsymbol{\theta}\in\mathbb{R}^{rp}}{\text{minimize}} \; t_2 \\ &\text{subject to} \; \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2^2 \leq t_2^2, \quad t_2 \geq 0, \\ &\qquad\qquad \|\boldsymbol{S}_2\boldsymbol{\theta}\|_2^2 \leq C_2, \end{aligned} \tag{4.16}$$

or, equivalently,

$$\begin{aligned} &\underset{t_2,\,\boldsymbol{\theta}}{\text{minimize}} \; t_2 \\ &\text{subject to} \; \|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{\theta}\|_2 \leq t_2, \\ &\qquad\qquad \|\boldsymbol{S}_2\boldsymbol{\theta}\|_2 \leq \sqrt{C_2}. \end{aligned} \tag{4.17}$$

As it can be observed, the optimization problem of Eq. (4.17) is compatible with the problem of Eq. (4.7), which is the standard form of CQP, with

$$\boldsymbol{c} = \left(1, \mathbf{0}_{rp+1}^T\right)^T, \boldsymbol{x} = \left(t_2, \boldsymbol{\theta}^T\right)^T, \boldsymbol{D}_1 = (\mathbf{0}_n, \boldsymbol{M}), \boldsymbol{d}_1 = \boldsymbol{b}, q_1 = 0, \boldsymbol{p}_1 = (1, 0, \ldots, 0)^T,$$

$\boldsymbol{D}_2 = \left(\mathbf{0}_{rp}, \boldsymbol{S}_2\right), \boldsymbol{d}_2 = \mathbf{0}_{rp}, \boldsymbol{p}_2 = \mathbf{0}_{rp+1}$ and $q_2 = -\sqrt{C_2}$.

The *dual problem* for problem Eq. (4.17) according to Eq. (4.9) is written as

$$\begin{aligned} &\text{maximize} \; \left(\boldsymbol{b}^T, 0\right)\boldsymbol{\tau}_1 + \left(\mathbf{0}_{rp}^T, -\sqrt{C_2}\right)\boldsymbol{\tau}_2 \\ &\text{subject to} \begin{pmatrix} \mathbf{0}_N^T & 1 \\ \boldsymbol{M}^T & \mathbf{0}_{rp+1} \end{pmatrix}\boldsymbol{\tau}_1 + \begin{pmatrix} \mathbf{0}_{rp}^T & 0 \\ \boldsymbol{S}_2^T & \mathbf{0}_{rp} \end{pmatrix}\boldsymbol{\tau}_2 = \begin{pmatrix} 1 \\ 0_{rp+1} \end{pmatrix}, \\ &\qquad \boldsymbol{\tau}_1 \in L^{N+1}, \boldsymbol{\tau}_2 \in L^{rp+1}. \end{aligned} \tag{4.18}$$

Moreover, $\left(t_2, \boldsymbol{\theta}, \boldsymbol{\chi}_2, \boldsymbol{\eta}_2, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2\right)$ is a *primal-dual optimal* C-LASSO solution if and only if the following constraints are satisfied:

$$
\begin{aligned}
&\boldsymbol{\chi}_2 = \begin{pmatrix} \mathbf{0}_N & \boldsymbol{M} \\ 1 & \mathbf{0}_{rp+1}^T \end{pmatrix} \begin{pmatrix} t_2 \\ \boldsymbol{\theta} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{b} \\ 0 \end{pmatrix}, \\
&\boldsymbol{\eta}_2 = \begin{pmatrix} \mathbf{0}_{rp} & \boldsymbol{S}_2 \\ 0 & \mathbf{0}_{rp}^T \end{pmatrix} \begin{pmatrix} t_2 \\ \boldsymbol{\theta} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{rp} \\ \sqrt{C_2} \end{pmatrix}, \\
&\begin{pmatrix} \mathbf{0}_N^T & 1 \\ \boldsymbol{M}^T & \mathbf{0}_{rp+1} \end{pmatrix} \boldsymbol{\tau}_1 + \begin{pmatrix} \mathbf{0}_{rp}^T & 0 \\ \boldsymbol{S}_2^{\mathrm{T}} & \mathbf{0}_{rp} \end{pmatrix} \boldsymbol{\tau}_2 = \begin{pmatrix} 1 \\ \mathbf{0}_{rp+1} \end{pmatrix}, \\
&\boldsymbol{\tau}_1^T \boldsymbol{\chi}_2 = 0, \quad \boldsymbol{\tau}_2^T \boldsymbol{\eta}_2 = 0, \\
&\boldsymbol{\tau}_1 \in L^{N+1}, \quad \boldsymbol{\tau}_2 \in L^{rp+1}, \quad \boldsymbol{\chi}_2 \in L^{N+1}, \quad \boldsymbol{\eta}_2 \in L^{rp+1}.
\end{aligned}
\tag{4.19}
$$

To solve convex optimization problems [4, 7], classical polynomial time algorithms can be applied. However, these algorithms have some disadvantages since they just use local information on the objective function and the constraints. For this reason, for solving "well-structured" convex optimization problems like conic quadratic problems, there are interior point methods [20, 24, 27] which were firstly introduced by Karmarkar for linear programming in 1984 [20]. Then, in the years, since these algorithms and software for linear programming have become quite developed, its extensions are used for more general classes of problems, such as convex quadratic programming, nonconvex and nonlinear problems over sets that can be characterized by self-concordant barrier functions. These algorithms are based on the given (primal) and the dual problem as well. They employ the structure of the problem in a global sense by allowing better complexity bounds and exhibit a much better practical performance.

Since in this present chapter, we expressed generalized additive spline regression problem for logistic regression as both L_1-regularization problem and a conic quadratic problem together with appropriate choice of a tolerance C_2 that it is should be the outcome of a careful learning process, with a model-free [17] or a model-based method [5], we become enabled to future research to exploit special problem for analytical and numerical purposes.

5 Concluding Remarks

In this chapter, we undertook an effort for a further introduction of modern continuous optimization into statistical learning and inverse problems. We presented an additive logistic regression model with B-spline and C-LASSO parameter estimation for them. Differently from estimation methods which employ penalty parameters with the need to regularly adjust them in the course of an algorithm, e.g., of Newton–Raphson type, a well-structured class of convex optimization method called

Conic Quadratic Programming offers an elegant framework of description analysis, a duality theory, an efficient algorithm and a reasonable complexity.

References

1. Aitchison, P. W. (1982). Generalized inverse matrices and their applications. *International Journal of Mathematical Education in Science and Technology, 13*(1), 99–109.
2. Aster, R. C., Borchers, B., & Thurber, C. H. (2013). *Parameter estimation and inverse problems*. New York: Academic Press.
3. Bagirov, A., Karmitsa, N., & Mäkelä, M. M. (2014). *Introduction to nonsmooth optimization: Theory*. Springer, Switzerland: Practice and Software.
4. Ben-Tal, A., & Nemirovski, A. (2001). Lectures on modern convex optimization: Analysis, algorithms and engineering applications. *MOS-SIAM Series on Optimization* (Philadelphia).
5. Blumschein, P., Hung, W., & Jonassen, D. (2009). *Model-based approaches to learning: Using systems models and simulations to improve understanding and problem solving in complex domains*. Rotterdam: Sense Publishers.
6. Bock, H. H., Sokolowski, A., & Jajuga, K. (2002). *Classification, clustering, and data analysis: Recent advances and applications*. New York: Springer.
7. Boyd, S., Vandenberghe, & L. Vandenberghe. (2004). *Convex optimization*. Cambridge University Press, Cambridge.
8. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical information theoretic approach*. New York: Springer.
9. Cox, M. G. (1982). Practical spline approximation. In: *Topics in Numerical Analysis. Lecture Notes in Mathematics*, vol. 630. Berlin: Springer.
10. De Boor, C. (2001). *Practical guide to splines*. New York: Springer.
11. Eggermont, P. P., Paul, P., & La Riccia, V. N. (2009). *Maximum penalized likelihood es timation, volume II: Regression*. New York: Springer.
12. Fletcher, R. (1987). *Practical methods of optimization*. New York: Wiley.
13. Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215–223.
14. Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations*. Baltimore: The John Hopkins Universi ty Press.
15. Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310.
16. Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.
17. Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The element of statistical learning*. New York: Springer.
18. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.
19. Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.
20. Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica, 4,* 373–395.
21. Kelley, C.T. (1995). Iterative methods for linear and nonlinear equations. *MOS-SIAM Series on Optimization* (Philadelphia).
22. Lobo, M. S., Vandenberghe, L., Boyd, S., & Lebret, H. (1998). Applications of second-order cone programming. *Linear Algebra and its Applications, 284,* 193–228.
23. McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman and Hall/CRC Mono graphs on Statistics and Applied Probability.
24. Nesterov, Yu., & Nemirovski, A. (1994). *Interior-point polynomial methods in convex programming*. Philadelphia: MOS-SIAM Series on Optimization.

25. Pringle, R. M., Rayner, A. A. (1971). *Generalized inverse matrices with applications to statistics*. NewYork: Hafner Publishing.
26. Quarteroni, A., Sacco, R., & Saleri, F. (2000). *Numerical mathematics*. New York: Springer.
27. Renegar, J. (1987). *A mathematical view of interior-point methods in convex optimization*. Philadelphia: MOS-SIAM Series on Optimization.
28. Schmidt, M., Fung, G., Rosales, R. (2009). Optimization methods for L1-regularization. UBC Technical Report TR-2009-19.
29. Shevade, S. K., & Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics, 19*(17), 2246–2253.
30. Shor, N. Z. (1985). *Minimization methods for non-differentiable functions*. Berlin: Springer.
31. Taylan, P., Weber, G. W., & Beck, A. (2007). New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. *Optimization, 56*(5–6), 675–698.
32. Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Statistical Methodology), 58,* 267–288.
33. Walker, S., & Page, C. (2001). Generalized ridge regression and a generalization of the CP statistics. *Journal of Applied Statistics, 28*(7), 911–922.
34. Weisberg, S. (2005). *Applied linear regression*. New Jersey: Wiley.

An Efficient Bundle-Like Algorithm for Data-Driven Multi-objective Bi-level Signal Design for Traffic Networks with Hazardous Material Transportation

Suh-Wen Chiou

1 Introduction

For most urban road networks, transportation of hazmat is of primary concern to decision makers whilst the growing complexity of signal-controlled road junctions has increased the risk of sudden disruption. Many model-based decision support systems (DSS) are intended to provide some kind of optimization of desired decision criteria. For instance, revenue management aiming at identifying optimal price levels of services by analyzing forecasts and current sales levels has been applied in Airlines with success. The demand chain management employs optimization models that incorporate uncertainty, product relationships, and stock levels to determine prices for products that downstream retailer may have. Even more, the product can be strategically priced by consumer demand management DSS by taking advantage of large databases, and high bandwidth networking to propagate the data in real time, and faster computers to enable solutions of large models.

A data-driven DSS has also been regarded as the most popular approach to tackle a network design problem [27–29]. In urban areas most travel delays incurred by road users are occurring at signal-controlled junctions at which traffic movements are interrupted regularly by alternating traffic lights. In urban area traffic control road network, planning to both cope with continuously growing travel demand and alleviate increasing traffic congestion at signalized junctions becomes one of the most challenging and important issues facing decision makers at various levels of management. Optimization of signal settings with certain travel demand in a road network has attracted many researchers over past decades [1–3, 6, 14, 25, 31, 32, 37]. For example, TRANSYT [34] is recognized as one of the most useful cases of computerized decision support systems in studying optimization of real-world fixed-time signal timings. As indicated in [1], traffic flows and travel times are strongly

S.-W. Chiou (✉)
Department of Information Management, National Dong Hwa University, Hualien County, Taiwan
e-mail: chiou@mail.ndhu.edu.tw; chiou@gms.ndhu.edu.tw

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_11

influenced by operations of signals. Ways of using mathematical programming to solve a constrained optimization problem of signal timings and equilibrium flow have been well researched [6, 8, 10, 36]. A bi-level programming technique [13, 15] traditionally has been considered a more appropriate tool in tackling signal setting problem [8, 36]. According to [9, 11], a traffic control manager with the leader at the upper level is supposed to have all information about road travelers' objective function and constraints, while at the lower level road travelers with the followers know nothing about the leader but the strategy announced by him. Until the leader announces his optimal strategy, the followers are not able to solve their optimizing problem responding to the leader's strategy.

For a signal-controlled road network, one direct approach usually adopted by government to regulate the risk associated with hazmat shipment is the prohibition on the use of certain roads by hazmat traffic [5, 18, 19]. A more amiable and flexible policy for regulators on effective road control can be achieved through appropriate network design [21, 23, 33]. While hazmat transport risks can be reduced by limiting access on urban city roads to hazmat traffic, carriers will incur increased costs due to limited routing available. Planning to both cope with mitigation of public risk and minimization of travel cost becomes one of the most challenging issues [16, 33]. Considering the rationale response of carriers [21] were the first ones to propose a bi-level network design model for hazmat transportation with single objective of risk minimization. Despite risk model taking the form of expected consequence of incidents is regarded as the most-widely used one [16, 33], the development of data-driven risk-averse models for low probability but high risk hazmat routing has attracted the attention of researchers only fairly recently [4, 17]. For example [17] introduced catastrophe-avoidance models for hazmat routing. [4] proposed an alternative using a game-theoretic approach to reduce the maximum risk along a route.

In this chapter, a data-driven bi-level decision support system (DBSS) with multi-objective for urban traffic signal design is developed to regulate the risk associated with hazmat transportation whilst the generalized travel costs of carriers can be effectively reduced. A data-driven min-max risk model is introduced for hazardous materials with low probability but high consequence. The maximum risk exposure of population adjacent to selected routes of hazmat and that over entire road network can be minimized. For a signal-controlled road network, the travel delays at downstream signal-controlled junctions are explicitly considered and evaluated using a traffic model [34]. Details can be referred to recent results in [7, 35]. The generalized travel cost for carriers of hazmat traffic can be minimized over the entire road network through better signal control policies. In this chapter, a data-driven bi-level network design model (DBM) model embedded with min-max risk model (BM) is established. The performance index for DBM can be taken as a weighted sum of maximum risk exposure and a linear combination of the rate of delay and the number of stops for each link. For a general bi-level problem, as mentioned in the literature [22, 24], only local solutions can be found due to the non-linearity of the constraints at the lower level problem. At the best of author's knowledge, there is very limited work in the literature to simultaneously tackle transport risk and generalized travel cost for

hazmat transportation via effective traffic signal control. In this chapter we intend to fill this gap. Due to non-convexity of the DBM, a bounding solution aiming to reduce relative gaps between iterations is developed.

The contributions made from this chapter can be briefly stated as follows. First, a DBSS with multi-objective using a bundle-like algorithm is presented to mitigate maximum risk exposure whilst total travel delay for hazmat transportation can be effectively reduced. A DBM with multi-objective is presented via effective traffic signal control. Second, BM is proposed to determine the maximum risk exposure for both carriers at the lower level problem and decision makers at the upper level problem. Due to non-convex structure of DBM, thirdly, a bundle-like algorithm is proposed to stabilize solutions of the problem and effectively solve the problem with reasonable computational efforts. Finally, numerical computations are performed using an example road network to demonstrate the superiority of proposed strategy. Comparisons are also made with other data-driven risk-averse models to evaluate the effectiveness of the proposed model. The results strongly indicate that DBSS becomes even more amiable to signal control policy makers since provides more flexible solutions whilst is more acceptable to hazmat carriers since explicitly takes account of travel delay at downstream signal-controlled junctions. Moreover, the trade-offs between maximum risk exposure and generalized travel costs are empirically investigated among different data-driven risk models with a variety of weights. As a result, DBSS consistently exhibits considerable advantage on mitigation of maximum risk whilst incurred less cost loss as compared to other data-driven risk-averse models.

The rest of the chapter is organized as follows. Section 2 introduces DBSS for signal control in hazmat transportation case study. In Sect. 3, a DBM with multi-objective is proposed. In Sect. 4, other data-driven risk-averse models for hazmat network design are presented. Section 5 gives a bundle-like algorithm for proposed DBM. Section 6 presents a data-driven bounding strategy and numerical computations are performed in Sect. 7. Conclusions for this chapter and extensions of the proposed approach to topics of interest are briefly summarized in Sect. 8.

2 A Data-Driven Bi-level Decision Support System (DBSS)

A data-driven bi-level decision support system using model-driven technique in [27–29] is developed in this section. In order to analyze the risk of hazmat transport to decision makers, a data-driven min-max risk model (BM) integrating with explicit signal delay in [7] is proposed in the first place. Notation used throughout this chapter is stated below.

2.1 Notation

Let $G(N, L)$ denote a directed road network, where N represents a set of fixed time signal controlled junctions and L represents a set of links denoted by $(i, j), \forall (i, j) \in L$. Each traffic stream approaching any junction is represented by its own link.

W	a set of origin-destination (OD) pairs.
R_w	a set of routes between OD pair w, $\forall w \in W$.
$T = [T_w]$	the matrix of travel demands for origin-destination pair w, $\forall w \in W$.
ζ	the reciprocal of the common cycle time.
$\zeta_{\min}, \zeta_{\max}$	the minimum and maximum reciprocal of the common cycle time.
$\theta = [\theta_{am}]$	the vector of starts of green for various links as proportions of cycle time where θ_{am} is start of next green for signal group a at junction m.
$\phi = [\phi_{am}]$	the vector of durations of green for various links as proportions of cycle time where ϕ_{am} is the duration of green for signal group a at junction m.
τ_{abm}	the clearance time between the end of green for signal group a and the start of green for incompatible signal group b at junction m.
$\Psi = (\zeta, \theta, \phi)$	the set of signal setting variables, respectively for the reciprocal of common cycle time, start and duration of greens.
$\lambda_{(i,j)}$	duration of effective green for link (i, j), $\forall (i, j) \in L$.
$\lambda_{\min}$	the minimum green.
$\Omega_m(a, b)$	collection of numbers 0 and 1 for each pair of incompatible signal groups at junction m; where $\Omega_m(a, b) = 0$ if the start of green for signal group a precedes that of b and $\Omega_m(a, b) = 1$, otherwise.
$D_{(i,j)}$	the rate of delay on link (i, j), $\forall (i, j) \in L$.
$S_{(i,j)}$	the number of stops per unit time on link (i, j), $\forall (i, j) \in L$.
W_D	weighting factor for rate of delay.
W_S	weighting factor for number of stops.
M_D	monetary factor associated with $D_{(i,j)}$.
M_S	monetary factor associated with $S_{(i,j)}$.
$\rho(i, j)$	maximum degree of saturation for link (i, j), $\forall (i, j) \in L$.
$S_{(i,j)}$	saturation flow on link (i, j), $\forall (i, j) \in L$.
$q_{(i,j)}$	incidental probability of accidental release of hazmat on link (i, j), $\forall (i, j) \in L$.
$r_{(i,j)}$	incidental consequence of accidental release of hazmat on link (i, j), $\forall (i, j) \in L$.
$f_{(i,j)}$	hazmat traffic flow on link (i, j), $\forall (i, j) \in L$.
h_k	hazmat traffic flow on route k between OD pairs, $\forall k \in R_w, w \in W$.
Λ	a link-route incidence matrix with entry $\Lambda^k_{(i,j)} = 1$ if route k uses link (i, j), and $\Lambda^k_{(i,j)} = 0$ otherwise, $\forall (i, j) \in L, \forall k \in R_w, w \in W$.
Γ	an OD-route incidence matrix with entry $\Gamma^w_k = 1$ if path k connects OD pair w, $\forall w \in W$, and $\Gamma^w_k = 0$ otherwise, $\forall k \in R_w, w \in W$.
$c_{(i,j)}$	the travel time on link (i, j), $\forall (i, j) \in L$.
$c^0_{(i,j)}$	the un-delayed travel time on link (i, j), $\forall (i, j) \in L$.

$d_{(i,j)}$	the average delay on link (i, j), $\forall(i, j) \in L$.
c_k	the travel time on route k.
σ	a converting factor from data-driven risk measure to monetary factor.

2.2 The Cost Minimum Route Optimization

For a signal-controlled road network, the travel time can be calculated as a sum of un-delayed travel time $c^0_{(i,j)}$ and the average delay $d_{(i,j)}$ incurred by hazmat traffic at downstream junction, i.e.

$$c_{(i,j)}(f, \Psi) = c^0_{(i,j)} + d_{(i,j)}(f, \Psi) \tag{1}$$

The cost minimum routes (CM) for hazmat carriers among pairs of OD $w, \forall w \in W$ can be found through a system optimum formulation.

$$\underset{f}{Min} \sum_{(i,j)\in L} f_{(i,j)} c_{(i,j)}(f, \Psi) \tag{2}$$

$$\text{subject to} \sum_{k\in R_w} h_k = T_w, \quad \forall_w \in W$$

$$f_{(i,j)} \sum_{w\in W} \sum_{k\in R_w} \Lambda^k_{(i,j)} h_k, \quad \forall(i, j) \in L$$

$$h_k \geq 0, \quad \forall k \in R_w, w \in W$$

Let Ω denote the feasible set for feasible traffic flow f in a following vector form, i.e. $\Omega\{f : f = \Lambda h, \Gamma h = T, h \geq 0\}$. For hazmat traffic flow f solving cost minimization (2), let Ω_{CM} denote a solution set for hazmat traffic flow.

2.3 A Linear Delay Risk Model

The expected consequence can be generalized as a following form: let $c^H_{(i,j)}$ denotes a public risk with probability $q_{(i,j)}$ of accidental release for hazmat to population exposure on link (i, j), we have a data-driven risk model as follows.

$$c^H_{(i,j)} = r_{(i,j)} q_{(i,j)} \tag{3}$$

A linear delay risk cost $c^G_{(i,j)}$ can be expressed as the following data-driven risk model.

$$c^G_{(i,j)}(f, \Psi) = c_{(i,j)}(f, \Psi)c^H_{(i,j)} \tag{4}$$

The corresponding route cost in relation to (3) and (4) can be expressed. Let c^H_k denote the route cost on route k, we have

$$c^H_k = \sum_{(i,j)\in L} c^H_{(i,j)}\Lambda^k_{(i,j)} \tag{5}$$

Let $c^G_k(f, \Psi)$ denote a linear delay risk route cost, we have

$$c^G_k(f, \Psi) = \sum_{(i,j)\in L} c^G_{(i,j)}(f, \Psi)\Lambda^k_{(i,j)} \tag{6}$$

A data-driven total travel cost can be expressed as follows.

$$TC^G(f, \Psi, q) = \sum_{(i,j)\in L} c_{(i,j)}(f, \Psi)f_{(i,j)}r_{(i,j)}q_{(i,j)} \tag{7}$$

Or in a vector form

$$TC^G(f, \Psi, q) = c(f, \Psi)frq \tag{8}$$

2.4 A Maximum Data-Driven Risk Model (MM)

Since the distribution of occurrence of probability q in data-driven travel cost (8) is not always available, the expected consequence of incident can be approximated by a maximum risk model. According to [17], a maximum risk link along a chosen route k can be identified as a following form: for every route k,

$$c^M_k = \underset{(i,j)\in k}{Max}\{c^H_{(i,j)}\} \tag{9}$$

Thus a least maximum risk model (MM) along a path k' can be expressed as follows.

$$\underset{k\in R_w, w\in W}{Min}\, c^M_k = \underset{k\in R_w, w\in W}{Min}\, \underset{(i,j)\in k}{Max}\{c^H_{(i,j)}\} \tag{10}$$

Find a route k' with a least risk such that for every route k, we have

$$k' = Arg\ \underset{k}{Min}\, c^M_k \tag{11}$$

2.5 A Maximum Data-Driven Risk with Mixed Routes (MM2) Model

Considering a mixed route strategy between a specified OD pair w [4], generalized a simplified data-driven min-max model as a following form. Let p_k denote path use probability for a trip w such that

$$1 = \sum_{k \in R_w} p_k \tag{12}$$

Let $p_{(i,j)}$ denote a link use probability such that for every link, by definition, we have

$$p_{(i,j)} = \sum_{k \in R_w} \Lambda_{(i,j)}^{k} p_k \tag{13}$$

The expected consequence in (3) with link use probability $p_{(i,j)}$ can be re-expressed as follows.

$$c_{(i,j)}^{H} = p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{14}$$

subject to (12) and (13). A maximum risk model $c^M(p)$ with a data-driven mixed-route selection probability $p_k, \forall k \in R_w$ between a specified OD pair w can be described as a following form: for any link use probability $p_{(i,j)}$ satisfying (12) and (13), we have

$$c^M(p) = \underset{q(i,j)}{Max} \sum_{(i,j) \in L} c_{(i,j)}^{H} = \underset{q(i,j)}{Max} \sum_{(i,j) \in L} p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{15}$$

$$\text{subject to } 1 = \sum_{(i,j) \in L} q_{(i,j)}$$

Therefore a data-driven maximum risk model (MM2) over entire road network can be re-expressed.

$$\underset{p_{(i,j)}}{Min}\, c^M(p) = \underset{p(i,j)}{Min}\, \underset{q(i,j)}{Max} \sum_{(i,j) \in L} p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{16}$$

$$\text{subject to } \sum_{(i,j) \in L} q_{(i,j)} = 1$$

$$\text{and } \sum_{k \in R_w} p_k = 1$$

$$\text{together with } p_{(i,j)} = \sum_{k \in R_w} \Lambda_{(i,j)}^{k} p_k$$

2.6 A Data-Driven Risk Model with Signal Delay

Assuming that the occurrence of accidental release for hazmat is barely to be known a priori, a data-driven min-max risk model is introduced according to (15).

$$\underset{q}{Max} \sum_{(i,j)\in L} f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{17}$$

$$\text{subject to} \quad 1 = \sum_{(i,j)\in L} q_{(i,j)}$$

According to (7), a data-driven risk model $TC^M(f, \Psi)$ taking account of signal delay $c(f, \Psi)$ can be described as a following maximum risk model.

$$TC^M(f, \Psi) = \underset{q(i,j)}{Max} \sum_{(i,j)\in L} c_{(i,j)}(f, \Psi) f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{18}$$

$$\text{subject to} \sum_{(i,j)\in L} q_{(i,j)} = 1$$

The occurrence of probability $q_{(i,j)}, \forall (i,j) \in L$ in (18) maximizing total data-driven risk for hazmat traffic over entire road network under signal setting Ψ can be determined.

$$q^M = Arg \underset{q}{Max} \sum_{(i,j)\in L} c_{(i,j)}(f, \Psi) f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{19}$$

$$\text{subject to} \sum_{(i,j)\in L} q_{(i,j)} = 1$$

Now we have a data-driven min-max risk model with signal delay in the following way:

$$TC^M(f, \Psi) = \sum_{(i,j)\in L} c_{(i,j)}(f, \Psi) f_{(i,j)} r_{(i,j)} q_{(i,j)}^M \tag{20}$$

Or in a vector form,

$$TC^M(f, \Psi) = c(f, \Psi) f \; rq^M \tag{21}$$

Therefore a data-driven min-max risk model (BM) with signal delay can be addressed as follows.

$$\underset{f}{Min}\,\underset{q}{Max} \sum_{(i,j)\in L} c_{(i,j)}(f,\Psi) f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{22}$$

$$\text{subject to} \sum_{k\in R_w} h_k = T_w, \quad \forall w \in W$$

$$f_{(i,j)} = \sum_{w\in W} \sum_{k\in R_w} \Lambda^k_{(i,j)} h_k, \quad \forall (i,j) \in L$$

$$h_k \geq 0, \quad \forall k \in R_w, w \in W$$

$$\sum_{(i,j)\in L} q_{(i,j)} = 1$$

According to (19) and (20), it implies

$$\underset{f}{Min} \sum_{(i,j)\in L} c_{(i,j)}(f,\Psi) f_{(i,j)} r_{(i,j)} q^M_{(i,j)} \tag{23}$$

$$\text{subject to} \sum_{k\in R_w} h_k = T_w, \quad \forall w \in W$$

$$f_{(i,j)} = \sum_{w\in W} \sum_{k\in R_w} \Lambda^k_{(i,j)} h_k, \forall (i,j) \in L$$

$$h_k \geq 0, \quad \forall k \in R_w, w \in W$$

The resulting hazmat traffic flow f thus can be determined with a most obnoxious incident probability q^M such that total risk is minimized in the worst case of accidental release to population exposure. That is

$$f(\Psi) = Arg\, \underset{f\in\Omega}{MinTC}^M(f', \Psi) \tag{24}$$

and let Ω_{BM} denote a solution set for data-driven hazmat flow satisfying (23).

2.7 A Variational Inequality for the Lower Level Problem

For a lower level problem (23), it can be generalized as a variational inequality if and only if the gradient of objective function with respect to hazmat traffic is available. It turns out that a marginal cost for a data-driven risk model in (21) can be expressed as follows.

$$\tilde{c}(f, \Psi) = \nabla_f TC^M(f, \Psi) = c(f, \Psi) rq^M + f \nabla c(f, \Psi) rq^M \tag{25}$$

The optimization problem (25) can be equivalently expressed as a following variational inequality: for every hazmat traffic flow $f' \in \Omega$, it is to find a hazmat traffic flow, $f \in \Omega$ with an obnoxious incident probability q^M such that

$$\tilde{c}(f, \Psi)(f' - f) \geq 0 \tag{26}$$

According to (25), for every hazmat traffic flow $f' \in \Omega_{BM}$, it implies

$$\tilde{c}(f', \Psi) = c(f', \Psi) rq^M + f' \nabla c(f', \Psi) rq^M \tag{27}$$

3 Data-Driven Bi-level Hazmat Network Design Model (DBM) with Multi-objective

A data-driven performance measure of signal-controlled network can be evaluated using a well-known traffic model TRANSYT [34]. According to [7], the performance measure in TRANSYT is represented as a sum for signal-controlled traffic streams (links) of a weighted linear combination of estimated rate of delay and number of stops per unit time.

3.1 A Delay-Minimizing Signal Setting Problem

Let P be a performance index for a signal-controlled road network with signal settings $\Psi = (\zeta, \theta, \phi)$ and traffic flow f. It can be generally expressed via function P_0 in terms of signal timings Ψ and network flow f in the following way.

$$\underset{\Psi, f}{Min}\, P = P_0(\Psi, f) \tag{28}$$

$$\text{subject to} \quad \Psi \in \Pi$$

The set Π defines the constraints of signal settings Ψ as follows. For a signal-controlled network, the cycle time constraint is expressed as

$$\zeta_{\min} \leq \zeta \leq \zeta_{\max} \tag{29}$$

For each signal controlled junction m, the phase a green time for all signal groups at junction m is expressed as

$$\lambda_{\min}\zeta \leq \phi_{am} \leq 1 \tag{30}$$

The link capacity for all links leading to junction m is expressed as

$$f_{(i,j)} \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)} \tag{31}$$

and the clearance time τ_{abm} for incompatible signal groups a and b at junction m is expressed as

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a, b) \tag{32}$$

The performance index P is taken to be a sum of a linear combination of the rate of delay and the number of stops per unit time for each link.

$$P = \sum_{(i,j)\in L} D_{(i,j)}(\Psi, f) W_D M_D + S_{(i,j)}(\Psi, f) W_S M_S \tag{33}$$

A delay-minimizing signal setting optimization problem can be expressed.

$$\underset{\Psi, f}{Min}\, P = \sum_{(i,j)\in L} D_{(i,j)}(\Psi, f) W_D M_D + S_{(i,j)}(\Psi, f) W_S M_S \tag{34}$$

$$\text{subject to} \quad \zeta_{\min} \leq \zeta \leq \zeta_{\max}$$

$$\lambda_{\min}\zeta \leq \phi_{am} \leq 1$$

$$f_{(i,j)} \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a, b)$$

3.2 A Data-Driven Risk-Averse Signal Setting Problem

Consider a data-driven risk-averse model in (17). A data-driven risk-averse model for signal setting optimization can be expressed as follows.

$$\underset{\Psi,f}{Min}\ \underset{q}{Max}\ P = \sum_{(i,j)\in L} c_{(i,j)}(f,\Psi) f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{35}$$

$$\text{subject to} \quad \zeta_{\min} \le \zeta \le \zeta_{\max}$$

$$\lambda_{\min}\zeta \le \phi_{am} \le 1$$

$$f_{(i,j)} \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \le \theta_{am} + \Omega_m(a,b)$$

$$\sum_{(i,j)\in L} q_{(i,j)} = 1$$

3.3 A DBM Signal Setting Optimization with Multi-objective

Considering two single objective signal-setting problems (34)–(35) with constraints (29)–(32), a DBM signal setting optimization with multi-objective can be expressed as a following weighted sum model.

$$\begin{aligned} P = & \sum_{(i,j)\in L} D_{(i,j)}(\Psi,f) W_D M_D + S_{(i,j)}(\Psi,f) W_S M_S \\ & + \sigma \underset{q}{Max} \sum_{(i,j)\in L} c_{(i,j)} f_{(i,j)} r_{(i,j)} q_{(i,j)} \end{aligned} \tag{36}$$

$$\text{subject to} \sum_{(i,j)\in L} q_{(i,j)} = 1$$

The occurrence of probability in (36) maximizing data-driven total risk over entire road network can be determined.

$$q^{BM} = Arg \underset{q}{Max} \sum_{(i,j)\in L} D_{(i,j)}(\Psi, f)W_D M_D + S_{(i,j)}(\Psi, f)W_S M_S + \sigma \sum_{(i,j)\in L} c_{(i,j)} f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{37}$$

$$\text{subject to} \sum_{(i,j)\in L} q_{(i,j)} = 1$$

Thus a data-driven weighted sum model can be re-expressed below.

$$P = \sum_{(i,j)\in L} D_{(i,j)}(\Psi, f)W_D M_D + S_{(i,j)}(\Psi, f)W_S M_S + \sigma \sum_{(i,j)\in L} c_{(i,j)} f_{(i,j)} r_{(i,j)} q^{BM}_{(i,j)} \tag{38}$$

For hazmat traffic in (24), i.e. $f \in \Omega_{BM}$, the data-driven performance index P in (28) can be re-expressed.

$$P_{BM} = P^c_{BM} + \sigma P^e_{BM} \tag{39}$$

A weighted sum of DBM can be expressed in the following way:

$$\underset{\Psi, f}{Min}\, P_{BM} = P^c_{BM} + \sigma P^e_{BM} \tag{40}$$

$$\text{subject to} \quad \zeta_{\min} \le \zeta \le \zeta_{\max}$$

$$\lambda_{\min}\zeta \le \phi_{am} \le 1$$

$$f_{(i,j)} \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \le \theta_{am} + \Omega_m(a, b)$$

$$\text{and} \quad f \in \Omega_{BM}$$

4 Alternative Data-Driven Risk-Averse Models

Considering the lower level problem in Sect. 2, in this section, various data-driven risk models are considered: a cost minimum route model (CM) in (2), a data-driven risk maximum model (MM) in (10) and a mixed-route data-driven risk maximum

model in (16). First, CM can be expressed below in which a risk-neutral flow is determined.

$$\underset{\Psi,f}{Min}\, P_{CM} = P_M^c + \sigma P_{CM}^e \tag{41}$$

$$\text{subject to} \quad \zeta_{\min} \le \zeta \le \zeta_{\max}$$

$$\lambda_{\min}\zeta \le \phi_{am} \le 1$$

$$f_{(i,j)} \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \le \theta_{am} + \Omega_m(a,b)$$

$$\text{and} \quad f \in \Omega_{CM}$$

Second, MM can be expressed below in which a risk-averse flow is defined. According to (9) and (10), a least maximum risk route k' can be determined such that for every route k, $k \in R_w$ between OD pair w, $w \in W$, we have

$$k' = Arg \underset{k}{Min} \underset{(i,j)\in L}{Max}\{c_{(i,j)}^H\} \tag{42}$$

Let $h_{k'} = T_w$, thus for every (i,j), $(i,j) \in L$, we have

$$f_{(i,j)} = \Lambda_{(i,j)}^{k'} h_{k'} \tag{43}$$

Let Ω_{MM} denote a solution set for hazmat flow satisfying (42) and (43). Then a data-driven MM model can be expressed.

$$\underset{\Psi,f}{Min}\, P_{MM} = P_{MM}^c + \sigma P_{MM}^e \tag{44}$$

$$\text{subject to} \quad \zeta_{\min} \le \zeta \le \zeta_{\max}$$

$$\lambda_{\min}\zeta \le \phi_{am} \le 1$$

$$f_{(i,j)} \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a, b)$$

$$\text{and} \quad f \in \Omega_{MM}$$

For a data-driven maximum risk model with mixed routes (MM2) given in [4] and specified in (16), a data-driven maximum risk hazmat flow can be determined.

$$\underset{f(i,j)}{Min}\ \underset{q(i,j)}{Max} \sum_{(i,j)\in L} f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{45}$$

$$\text{subject to} \quad 1 = \sum_{(i,j)\in L} q_{(i,j)}$$

$$\sum_{k\in R_w} h_k = T_w, \forall w \in W$$

$$\text{and} \quad f_{(i,j)} = \sum_{w\in W} \sum_{k\in R_w} \Lambda^k_{(i,j)} h_k$$

Let Ω_{MM2} denote a solution set for hazmat flow satisfying (45). Then MM2 can be expressed in which a data-driven risk-averse flow is determined by (45).

$$\underset{\Psi, f}{Min}\, P_{MM2} = P^c_{MM2} + \sigma P^e_{MM2} \tag{46}$$

$$\text{subject to} \quad \zeta_{\min} \leq \zeta \leq \zeta_{\max}$$

$$\lambda_{\min}\zeta \leq \phi_{am} \leq 1$$

$$f_{(i,j)} \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a, b)$$

$$\text{and} \quad f \in \Omega_{MM2}$$

5 A Bundle-Like Algorithm for DBM with Multi-objective

According to recent developments in [26], the perturbation ∇f of hazmat traffic flow in (26) with respect to a change Δ in signal settings Ψ can thus be determined by solving a following linearized variational inequality. Let $\Delta = \hat{\Psi} - \Psi$, introduce

$$\Omega(\Delta) = \{\nabla f(\Psi) : \nabla f = \Lambda(\Delta h), \Gamma(\Delta h) = 0, \exists \Delta h \in K\} \tag{47}$$

and

$$K = \left\{ \Delta h : \begin{array}{l} (i)\Delta h_k \quad free, \quad if \quad h_k^* > 0, \\ (ii)\Delta h_k \geq 0, if \quad h_k^* = 0, \tilde{C}_k = \pi_w, \ \forall k \in R_w, \forall w \in W \\ (iii)\Delta h_k = 0, if \quad h_k^* = 0, \tilde{C}_k > \pi_w \end{array} \right\} \tag{48}$$

In (48), $\tilde{C}_k$ denotes a marginal cost on route k and π_w denotes a minimum cost for trip w. For every flow perturbation in set (47), i.e. $f' \in \Omega(\Delta)$, a directional derivative ∇f along a direction Δ can be determined such that

$$(\nabla_\Psi \tilde{c}(f, \Psi^*)\Delta + \nabla_f \tilde{c}(f, \Psi^*)\nabla f)(f' - \nabla f) \geq 0 \tag{49}$$

The gradients $\nabla_\Psi \tilde{c}(f, \Psi^*)$ and $\nabla_f \tilde{c}(f, \Psi^*)$ in (49) are evaluated at Ψ^* when perturbations in Δ are specified. For the objective function in (40), the generalized gradient for data-driven performance measure can be calculated.

$$\begin{aligned} \nabla_\Psi P = &(\nabla_\Psi D(\Psi, f) + \nabla_f D(\Psi, f) f'(\Psi)) W_D M_D \\ &+ (\nabla_\Psi S(\Psi, f) + \nabla_f S(\Psi, f) f'(\Psi)) W_S M_S + \sigma(c' f + f') r q^{BM}) \end{aligned} \tag{50}$$

According to directional derivatives by (49), a data-driven bi-level signal setting optimization (40) with multi-objective can now be reduced into a single-level optimization problem.

$$\underset{\Psi}{Min}\, P_1 = \sum_{(i,j)\in L} D_{(i,j)} W_D M_D + s_{(i,j)} W_S M_S + \sigma \sum_{(i,j)\in L} c_{(i,j)} f_{(i,j)} r_{(i,j)} q_{(i,j)}^{BM} \tag{51}$$

$$\text{subject to } \zeta_{\min} \leq \zeta \leq \zeta_{\max}$$

$$\lambda_{\min} \zeta \leq \phi_{am} \leq 1$$

$$f_{(i,j)}(\Psi) \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}$$

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a, b)$$

The generalized gradients for problem P_1 with respect to signal settings Ψ can be obtained as follows.

$$\nabla_\Psi P(\Psi) = \nabla_\Psi P_1 \tag{52}$$

For a signal setting $\Psi^{(k)}$, the generalized gradient of performance $P_1(\Psi)$ can be expressed whenever a sub-gradient $z \in \partial P_1(\Psi)$ exists according to Rademacher's theorem in [12]. Let co denote a convex hull, it implies

$$\partial_\Psi P_1(\Psi^*) = co\left\{\lim_{k\to\infty} \nabla_\Psi P_1(\Psi^{(k)}) : \Psi^{(k)} \to \Psi^*, \nabla_\Psi P_1(\Psi^{(k)}) exists\right\} \tag{53}$$

In (53), the generalized gradient of $P_1(\Psi^{(k)})$ can be obtained according to (50) and (52), i.e.

$$\begin{aligned}\nabla_\Psi P_1(\Psi^{(k)}) &= (\nabla_\Psi D(\Psi^{(k)}, f(\Psi^{(k)})) + \nabla_f D(\Psi^{(k)}, f(\Psi^{(k)}))f'(\Psi^{(k)}))W_D M_D \\ &+(\nabla_\Psi S(\Psi^{(k)}, f(\Psi^{(k)})) + \nabla_f S(\Psi^{(k)}, f(\Psi^{(k)}))f'(\Psi^{(k)}))W_S M_S + \sigma(c'_{(i,j)}f + cf'_{(i,j)})rq^{BM}\end{aligned} \tag{54}$$

For a direction $\Delta^{(k)} = \Psi - \Psi^{(k)}$ in the neighborhood of signal setting $\Psi^{(k)}$, for any sub-gradient $z \in \partial P_1(\Psi^{(k)})$ the directional derivative $DP_1(\Psi^{(k)}; \Delta^{(k)})$ can be calculated as follows.

$$DP_1(\Psi^{(k)}; \Delta^{(k)}) = z\Delta^{(k)} \tag{55}$$

A tentative solution $\hat{\Psi}^{(k)}$ for (51) can be determined by a following bundle-like algorithm.

$$\hat{\Psi}^{(k)} = Arg \underset{\Psi\in\Pi}{Min}\ \underset{1\leq i\leq k}{Max}\ z^{(i)}(\Psi - \Psi^{(k)}) \tag{56}$$

A lower bound for (51) is denoted by P_l, and can be expressed as $P_l^{(k)} = P_1(\hat{\Psi}^{(k)})$. A new signal setting $\Psi^{(k+1)}$ is determined such that the performance index in (51) can be effectively reduced, that is, let $\Delta^{(k)} = \hat{\Psi}^{(k)} - \Psi^{(k)}$

$$P_1(\Psi^{(k+1)}) = P_1(\Psi^{(k)} + \alpha^{(k)}\Delta^{(k)}) \leq P_1(\Psi^{(k)} + \xi\Delta^{(k)}) < P_1(\Psi^{(k)}) \tag{57}$$

A practical Armijo line search rule can be conveniently utilized to determine the value of $\alpha^{(k)}$ in greatly reducing computational efforts for successive evaluations of the performance function $P_1(\Psi^{(k)})$. Let $z^{(k)} \in \partial_\Psi P_1(\Psi^{(k)})$, for chosen scalars β, l and γ with $0 < \beta < 1$, and $0 < \gamma < 1$, we set $\alpha^{(k)} = \beta^{m^{(k)}} l$ where $m^{(k)}$ is the first non-negative integer m for which

$$P_1(\Psi^{(k)} + \beta^m l\Delta^{(k)}) - P_1(\Psi^{(k)}) \leq \gamma\beta^m lz^{(k)}\Delta^{(k)} \tag{58}$$

A new signal setting for iteration $k+1$ can be expressed as follows.

$$\Psi^{(k+1)} = \Psi^{(k)} + \alpha^{(k)}\Delta^{(k)} \tag{59}$$

The upper bound for (51) is denoted by P_u and can be expressed as $P_u^{(k+1)} = P_1(\Psi^{(k+1)})$. Let $\delta^{(k)}$ denote a relative difference between bounds such that

$$\delta^{(k)} = P_u^{(k)} - P_l^{(kk)} \tag{60}$$

A data-driven bounding strategy for (51) is developed such that computation of signal settings for hazmat traffic flow can proceed until the relative difference $\delta^{(k)}$ between the upper bound $P_u^{(k)}$ and lower bound $P_l^{(k)}$ vanishes and $\Psi^{(k)}$ is the solution of (51).

6 A Data-Driven Bounding Strategy for DBM

For (51), the signal settings minimizing total traffic queues and maximum risk can be determined by a following data-driven bounding strategy. According to (58), the corresponding performance measure in (51) can be continuously improved from iteration to iteration when optimal signal settings are updated in (59) until the relative gap between bounds of performance function vanishes.

Step 1. Start with initial signal-setting $\Psi^{(k)}$ and set index $k=1$. Set the stopping threshold ε, $\varepsilon \geq 0$.
Step 2. Solve the lower level problem with signal-setting $\Psi^{(k)}$.

Step 2-1. Solve a least BM model with signal setting delay in (22).
Step 2-2. For a most obnoxious incidents with probability q^M, find a data-driven maxi-sum risk flow f via (24).
Step 2-3. Characterize the marginal cost for carriers via (25).

Step 3. Find directional derivatives via solving a linearized variational inequality (49) with (47) and (48).
Step 4. Determine the generalized gradient in (50).
Step 5. Solve DBM (51):

Step 5-1. Find a tentative signal setting $\hat{\Psi}^{(k)}$ via (56).
Step 5-2. Calculate a tentative direction $\Delta^{(k)} = \hat{\Psi}^{(k)} - \Psi^{(k)}$.
Step 5-3. Compute $DP_1(\Psi^{(k)}; \Delta^{(k)})$. If $DP_1(\Psi^{(k)}; \Delta^{(k)}) > -\varepsilon$, go to Step 6-3. Otherwise go to Step 5-4.
Step 5-4. Set a lower bound $P_l^{(k)} = \hat{P}_l^{(k)}(\hat{\Psi}^{(k)})$.

Advanced Regression Models: Least Squares, Nonlinear, Poisson and Binary Logistics Regression Using R

William P. Fox and Jesse Hammond

1 Introduction

Analysis in data science and digital business requires analysis of the data and in many cases the use of regression techniques. This chapter discusses some simple regression and advanced regression techniques that have bene used often in the analysis of data for business, industry, and government. Regression is not a one method fits all approach. Regression takes good approaches and common sense to complement the mathematical and statistical approaches used. We also discuss methods to check for model adequacy after the regression model is found. We also believe because of the popularity of R that we would illustrate this chapter using R and at the end of this chapter we provide the commands that were used in our examples.

Often we might want to model the data in order to make predictions or explain within the domain of the data, Besides the models, we provide insights into the adequacy of the model through various approaches including regression ANOVA output, residual plots, and percent relative error.

In general, we suggest using the following steps in regression analysis.

Step 1. Enter the data (x, y) and obtain a scatterplot of the data and note the trends.
Step 2. If necessary, transform the data into "y" and "x" components.
Step 3. Build or compute the regression equation. Obtain all the output. Interpret the ANOVA output for R^2, *F-test*, and *P*-values for coefficients.
Step 4. Plot the regression function and the data to obtain a visual fit.
Step 5. Compute the predictions, the residuals, percent relative error as described later.
Step 6. Insure the predictive results passes the common sense test.
Step 7. Plot the residual versus prediction to determine model adequacy.

W. P. Fox (✉) · J. Hammond
Naval Postgraduate School, Monterey, USA
e-mail: wpfox@nps.edu

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_12

We present several methods to check for model adequacy. First, we suggest your predictions pass the "common sense" test. If not, return to your regression model as we shown with our exponential decay model in Sect. 3. The residual plot is also very revealing. Figure 1 shows possible residual plot results where only random patterns indicate model adequacy from the residual plot perspective. Linear, curve, or fanning trend indicate a problem in the regression model. Afifi and Azen [1] have a good and useful discussion on corrective action based upon trends found. Percent relative error also provides information about how well the model approximates the original values and it provides insights into where the model fits well and where it might not fit well. We define percent relative error with Eq. (1),

$$\%RE = \frac{100|y_a - y_p|}{y_a} \tag{1}$$

2 Introducing Linear Regression

2.1 Correlation of Spring Data

First, let's define correlation. Correlation, ρ, measures the linearity between the data sets X and Y. Mathematically correlation is defined as follows:

The correlation coefficient, Eq. (2), between X and Y, denoted as ρ_{xy}, is

$$\rho_{xy} = \frac{COV(X,Y)}{\sigma_x \sigma_y} = \frac{E[XY] - \mu_x \mu_y}{\sigma_x \sigma_y}. \tag{2}$$

The values of correlation range from -1 to $+1$. The value of -1 corresponds to a perfect line with a negative slope and a value of $+1$ corresponds to a perfect line with a positive slope. A value of 0 indicates that that there is no linear relationship.

We present two rules of thumb for correlation from the literature. First, from Devore [2], for math, science, and engineering data we have the following:

$0.8 < |\rho| \leq 1.0$ Strong linear relationship
$0.5 < |\rho| \leq 0.8$ Moderate linear relationship
$|\rho| \leq 0.5$ Weak linear relationship

According to Johnson [8] for non-math, non-science and non-engineering data we find a more liberal interpretation of ρ:

$0.5 < |\rho| \leq 1.0$ Strong linear relationship
$0.5 < |\rho| \leq 0.3$ Moderate linear relationship
$0.1 < |\rho| \leq 0.3$ Weak linear relationship
$|\rho| \leq 0.1$ No linear relationship

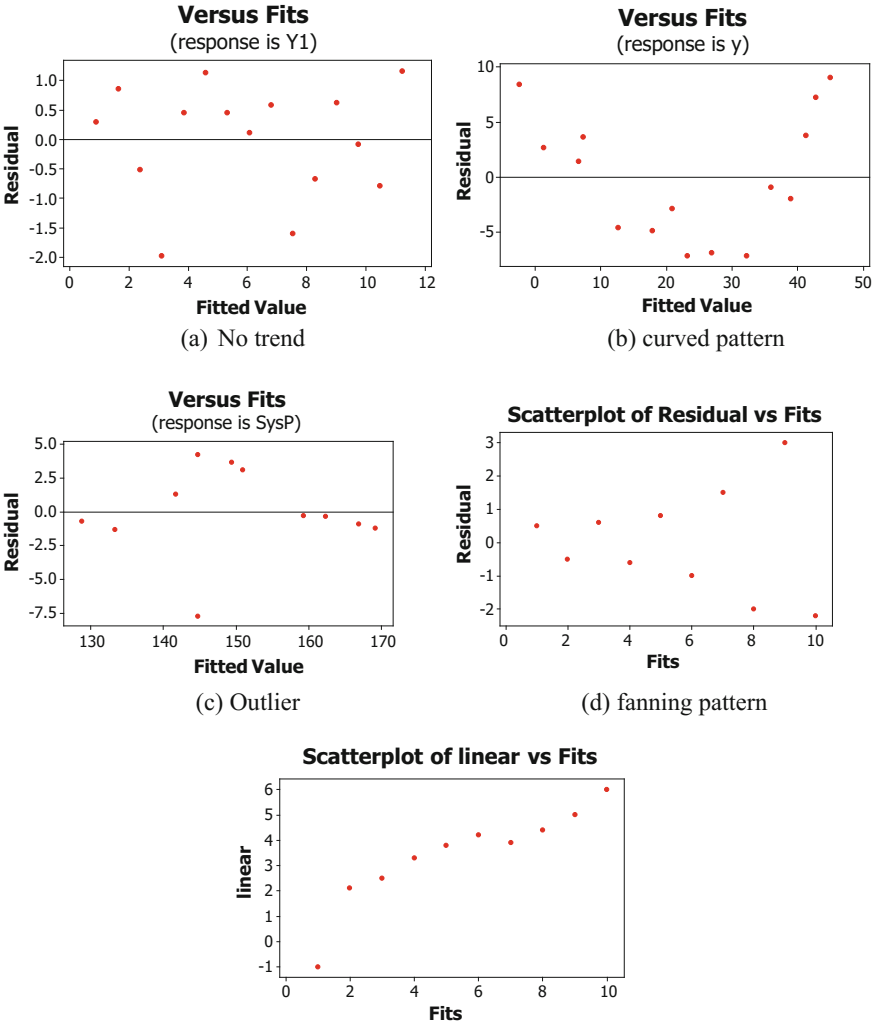


Fig. 1 Patterns for residuals **a** no pattern, **b** curved pattern, **c** outliers, **d** fanning pattern, **e** linear trend

Further, in our modeling efforts we emphasize the interpretation of $|\rho| \approx 0$. This can be interpreted as either no linear relationship or the existence of a nonlinear relationship. Most students and many researchers fail to pick up on the importance of the nonlinear relationship aspect of the interpretation.

Calculating correlation between two (or more) variables in *R* is simple. After loading in the spring data set as an object into *R*'s workspace, we can first visualize the data in tabular format. This lets us be sure that the data is in the proper format

and that there are no oddities (missing values, characters entered instead of numbers) that would cause problems.

Using either rule of thumb the correlation coefficient, $|\rho| = 0.999272$, indicates a strong linear relationship. We obtain this value, look at Fig. 1, and we see an excellent linear relationship with a positive correlations very close to 1.

	x	*y*
1	*50*	*0.1000*
2	*100*	*0.1875*
3	*150*	*0.2750*
4	*200*	*0.3250*
5	*250*	*0.4375*
6	*300*	*0.4875*
7	*350*	*0.5675*
8	*400*	*0.6500*
9	*450*	*0.7250*
10	*500*	*0.8000*
11	*550*	*0.8750*

To estimate the correlation between the two columns in this data set, we simply use the *cor()* command in *R* on the data table:

```
## Calculate and print correlation matrix
print(cor(spring_data))
```

```
##          x         y
## x 1.0000000 0.9992718
## y 0.9992718 1.0000000
```

The data's correlation coefficient is 0.9992718 that is very close to 1. Visualizing the data makes this relationship easy to see and we would expect to see a linear relationship with a positive slope as shown in Fig. 2.

2.2 Linear Regression of Spring Data

Fitting an ordinary least-squares (OLS) model with form $y\,x + \epsilon$ to the spring data in R is quite simple. Using the *lm()* command (short for "linear model") fits the linear model and saves the result as another object in *R*'s workspace.

```
## Fit OLS model to the data
spring_model <- lm(
  y ~ x
  , data = spring_data
  )
```

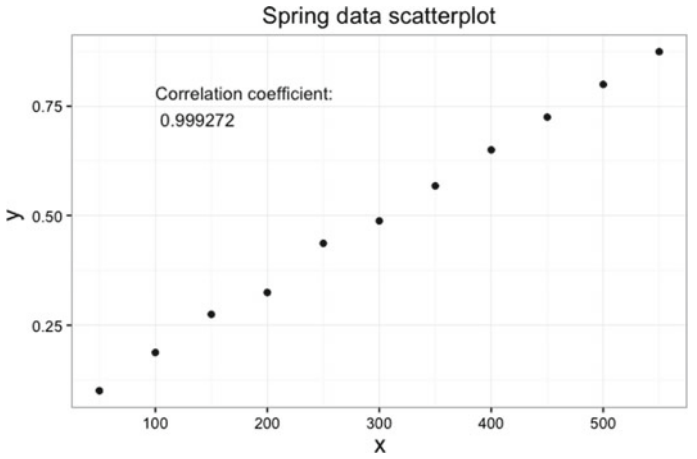


Fig. 2 Plot of spring data with correlation value

We can then perform operations on this object to produce tables presenting coefficient estimates and a range of diagnostic statistics to evaluate how well the model fits the data provided.

	Estimate	Std. error	t value	Pr(>|t|)
x	0.001537	1.957e−05	78.57	4.437e−14
(Intercept)	0.03245	0.006635	4.891	0.0008579

Fitting linear model: $y \sim x$

Observations	Residual std. error	R^2	Adjusted R^2
11	0.01026	0.9985	0.9984

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	0.6499	0.6499	6173	4.437e−14
Residuals	9	0.0009475	0.0001053	NA	NA

We visualize this estimated relationship by overlaying the fitted line to the spring data plot. This plot shows that the trend line estimated by the linear model fits the data quite well as shown in Fig. 3. The relationship between R^2 and ρ is that $R^2 = (\rho)^2$.

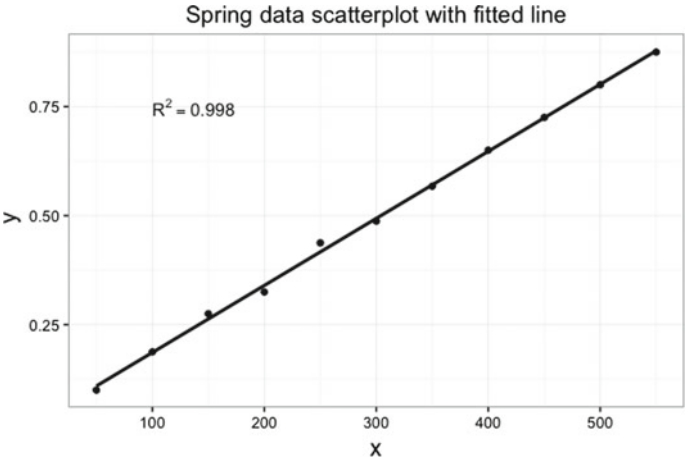


Fig. 3 Regression plot of spring data

2.3 Linear Regression of Philippines SIGACTS

Linear regression is not the answer to all analysis. The example below uses the same *lm()* command in R to estimate the relationship between literacy rates and violent events in the Philippines. Although the model produces a linear trend, both the fit statistics and visual inspection suggest that a linear model may not be the best choice.

```
## Fit OLS model to the data
sigacts_model_ols <- lm(
  sigacts_2008 ~ literacy
  , data = sigacts_data)
```

	Estimate	Std. error	t value	Pr(>|t|)
literacy	−1.145	0.4502	−2.543	0.01297
(Intercept)	113	37.99	2.975	0.003903

Fitting linear model: sigacts_2008 ~ literacy

Observations	Residual std. error	R^2	Adjusted R^2
80	25.77	0.07656	0.06472

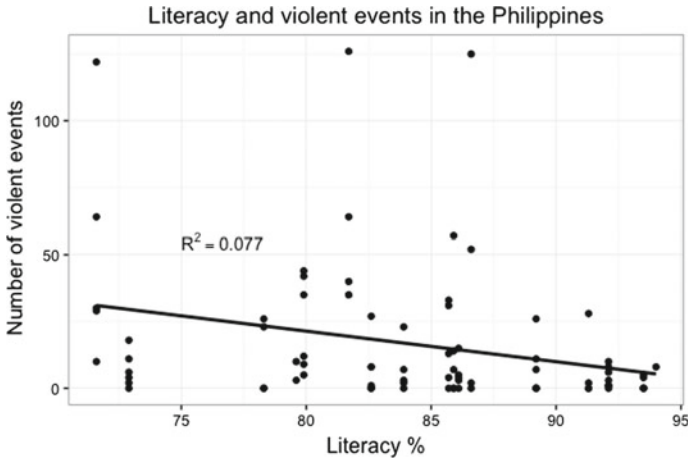


Fig. 4 Regression with literacy and violent events data

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Literacy	1	4295	4295	6.467	0.01297
Residuals	78	51,805	664.2	NA	NA

As seen in this example between literacy and violent events the linear regression model, Fig. 4, is not helpful. We will return to this example later in this chapter.

3 Exponential Decay Modeling

3.1 Introducing Hospital Recovery Data

There often arise cases where a linear model may not be appropriate. One potential case analyzed here is the relationship hospital stay and patient recovery. We begin by visualizing the data in R:

```
print(recovery_data)
## # A tibble: 15 × 2
##        T     Y
##    <int> <int>
## 1      2    54
## 2      5    50
## 3      7    45
## 4     10    37
## 5     14    35
## 6     19    25
## 7     26    20
## 8     31    16
## 9     34    18
## 10    38    13
## 11    45     8
## 12    52    11
## 13    53     8
## 14    60     4
## 15    65     6
```

Printing the table of recovery data shows that once again, the structure of the data is amenable to statistical analysis—there are no missing values and both columns contain valid numbers (integers). We have two columns, T (number of days in the hospital) and Y (estimated recovery index) and we want to generate a model that predicts how well a patient will recover as a function of the time they spend in the hospital. Using the *cor()* command retrieves an initial correlation coefficient of -0.941.

```
## Calculate and print correlation matrix
print(cor(recovery_data))

##           T         Y
## T  1.0000000 -0.9410528
## Y -0.9410528  1.0000000
```

Once again, creating a scatter plot, Fig. 5, of the data helps us visualize how closely the estimated correlation value matches the overall trend in the data.

In this example we will show linear regression, polynomial regression, and then exponential regression in order to obtain a useful model.

3.2 Linear Regression of Hospital Recovery Data

It definitely appears that there is a strong negative relationship: the longer a patient spends in the hospital, the lower their recovery index tends to be. Next, we fit an OLS model to the data to estimate the magnitude of the linear relationship.

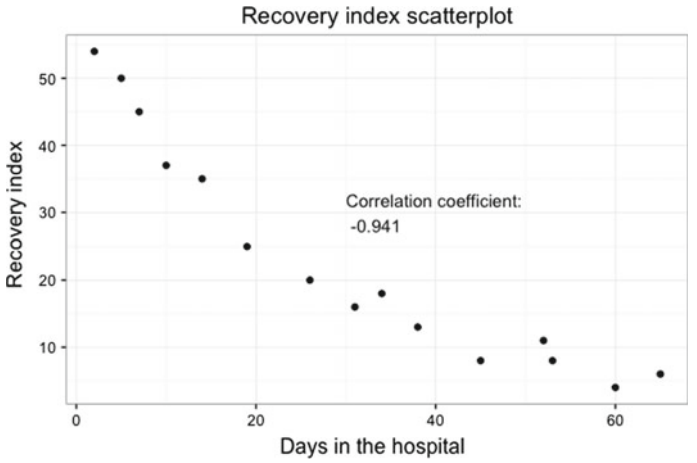


Fig. 5 Scatterplot of days in the hospital and recovery index

```
## Fit OLS model to the data
recovery_model <- lm(
  Y ~ T
  , data = recovery_data
  )
```

	Estimate	Std. error	t value	Pr(>|t|)
T	−0.7525	0.07502	−10.03	1.736e−07
(Intercept)	46.46	2.762	16.82	3.335e−10

Fitting linear model: Y ~ T

Observations	Residual std. error	R^2	Adjusted R^2
15	5.891	0.8856	0.8768

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
T	1	3492	3492	100.6	1.736e−07
Residuals	13	451.2	34.71	NA	NA

OLS modeling shows that there is a negative and statistically significant relationship between time spent in the hospital and patient recovery index. However, ordinary

least-squares regression may not be the best choice in this case for two reasons. First, we are dealing with real-world data: a model that can produce (for example) negative estimates of recovery index is not applicable to the underlying concepts our model is dealing with. Second, the assumption of OLS, like all linear models, is that the magnitude of the relationship between input and output variables stays constant over the entire range of values in the data. However, visualizing the data suggests that this assumption may not hold—in fact, it appears that the magnitude of the relationship is very high for low values of T and decays somewhat for patients who spend more days in the hospital.

To test for this phenomenon we examine the residuals of the linear model. Residuals analysis can provide quick visual feedback about model fit and whether the relationships estimated hold over the full range of the data. We calculate residuals as the difference between observed values Y and estimated values Y^*, or $Y_i - Y_i^*$. We then normalize residuals as percent relative error between the observed and estimated values, which helps us compare how well the model predicts each individual observation in the data set:

```
## # A tibble: 15 × 6
##       T     Y index predicted  residuals pct_relative_error
##   <int> <int> <int>     <dbl>      <dbl>              <dbl>
## 1     2    54     1 44.955397  9.0446035          16.749266
## 2     5    50     2 42.697873  7.3021275          14.604255
## 3     7    45     3 41.192857  3.8071435           8.460319
## 4    10    37     4 38.935333 -1.9353325          -5.230628
## 5    14    35     5 35.925301 -0.9253005          -2.643716
## 6    19    25     6 32.162761 -7.1627605          -28.651042
## 7    26    20     7 26.895205 -6.8952045          -34.476023
## 8    31    16     8 23.132665 -7.1326645          -44.579153
## 9    34    18     9 20.875141 -2.8751405          -15.973003
## 10   38    13    10 17.865109 -4.8651085           -37.423912
## 11   45     8    11 12.597553 -4.5975525          -57.469407
## 12   52    11    12  7.329997  3.6700035           33.363668
## 13   53     8    13  6.577489  1.4225115          17.781393
## 14   60     4    14  1.309933  2.6900675          67.251686
## 15   65     6    15 -2.452607  8.4526075          140.876791
```

These data can also be plotted to visualize how well the model fits over the range of our input variable.

The residuals plotted, Fig. 6, show a curvilinear pattern, decreasing and then increasing in magnitude over the range of the input variable. This means that we can likely improve the fit of the model by allowing for non-linear effects. Furthermore, the current model can make predictions that are substantively nonsensical, even if they were statistically valid. For example, our model predicts that after 100 days in the hospital, a patient's estimated recovery index value would be -29.79. This has no common sense, as the recovery index variable is always positive in the real world. By allowing for non-linear terms, we can also guard against these types of nonsense predictions.

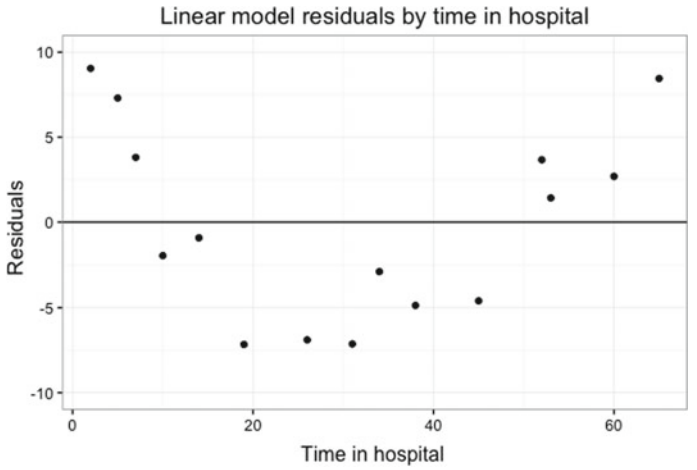


Fig. 6 Residual plot for linear model

3.3 Quadratic Regression of Hospital Recovery Data

Including a quadratic term modifies the model formula: $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ Fitting this model to the data produces separate estimates of the effect of T itself as well as the effect of T^2, the quadratic term.

```
## Generate model
recovery_model2 <- lm(Y ~ T + I(T^2), data = recovery_data)
```

	Estimate	Std. error	t value	Pr(>|t|)
T	−1.71	0.1248	−13.7	1.087e−08
I(T^2)	0.01481	0.001868	7.927	4.127e−06
(Intercept)	55.82	1.649	33.85	2.811e−13

Fitting linear model: Y ~ T + I(T^2)

Observations	Residual std. error	R^2	Adjusted R^2
15	2.455	0.9817	0.9786

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
T	1	3492	3492	579.3	1.59e−11
I(T^2)	1	378.9	378.9	62.84	4.127e−06
Residuals	12	72.34	6.029	NA	NA

Including the quadratic term improves model fit as measured by R^2 from 0.88 to 0.98—a sizable increase. To assess whether this new input variable deals with the curvilinear trend we saw in the residuals from the first model, we calculate and visualize the residuals from the quadratic model:

```
## # A tibble: 15 × 6
##        T     Y index predicted  residuals pct_relative_error
##    <int> <int> <int>     <dbl>      <dbl>              <dbl>
## 1      2    54     1 52.460836  1.5391644          2.8503045
## 2      5    50     2 47.640993  2.3590072          4.7180144
## 3      7    45     3 44.575834  0.4241663          0.9425917
## 4     10    37     4 40.200199 -3.2001992         -8.6491871
## 5     14    35     5 34.780614  0.2193857          0.6268164
## 6     19    25     6 28.672445 -3.6724455        -14.6897820
## 7     26    20     7 21.364792 -1.3647924         -6.8239618
## 8     31    16     8 17.033457 -1.0334567         -6.4591042
## 9     34    18     9 14.790022  3.2099781         17.8332119
## 10    38    13    10 12.213370  0.7866302          6.0510012
## 11    45     8    11  8.844363 -0.8443634        -10.5545422
## 12    52    11    12  6.926437  4.0735627         37.0323886
## 13    53     8    13  6.770903  1.2290967         15.3637082
## 14    60     4    14  6.511355 -2.5113548        -62.7838691
## 15    65     6    15  7.214379 -1.2143795        -20.2396576
```

Visually, Fig. 7, evaluating the residuals from the quadratic model shows that the trend has disappeared. This means that we can assume the same relationship holds whether $T = 1$ or $T = 100$. However, we are still not sure if the model produces numerical estimates that pass the common-sense test. The simplest way to assess this is to generate predicted values of the recovery index variable using the quadratic model, and plot them to see if they make sense.

To generate predicted values in R, we can pass the quadratic model object to the *predict()* function along with a set of hypothetical input values. In other words, we can ask the model what the recovery index would look like for a set of hypothetical patients who spend anywhere from zero to 120 days in the hospital.

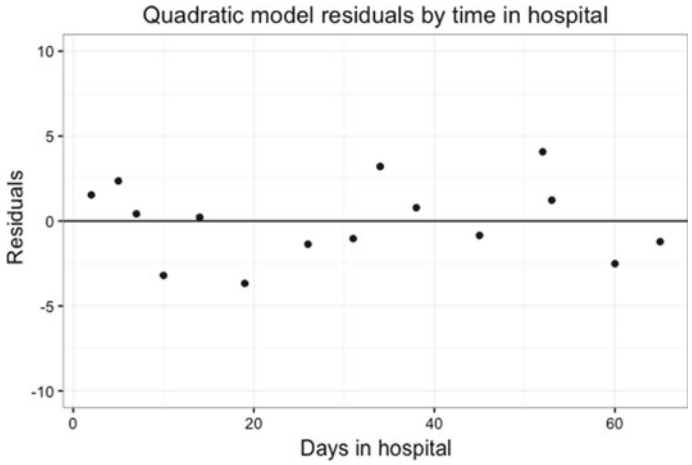


Fig. 7 Residual plot for polynomial regression model

```
## Create a set of hypothetical patient observations with days in the hospital from 1 to 120
patient_days = tibble(T = 1:120)

## Feed the new data to the model to generate predicted recovery index values
predicted_values = predict(
  recovery_model2
  , newdata = patient_days
  )
```

We can then plot these estimates to quickly gauge whether they pass the common-sense test for real-world predictive value as shown in Fig. 7.

```
## # A tibble: 5 × 2
##       T predicted
##   <int>     <dbl>
## 1     1  54.12668
## 2     2  52.46084
## 3     3  50.82461
## 4     4  49.21799
## 5     5  47.64099
```

The predicted values curve up toward infinity, Fig. 8; clearly this is a problem. The quadratic term we included in the model leads to unrealistic estimates of recovery index at larger values of T. Not only is this unacceptable for the context of our model, but it is unrealistic on its face. After all, we understand that people generally spend long periods in the hospital for serious or life-threatening conditions such as severe disease or major bodily injury. As such, we can assess that someone who spends six months in the hospital probably should not have a higher recovery index than someone who was only hospitalized for a day or two.

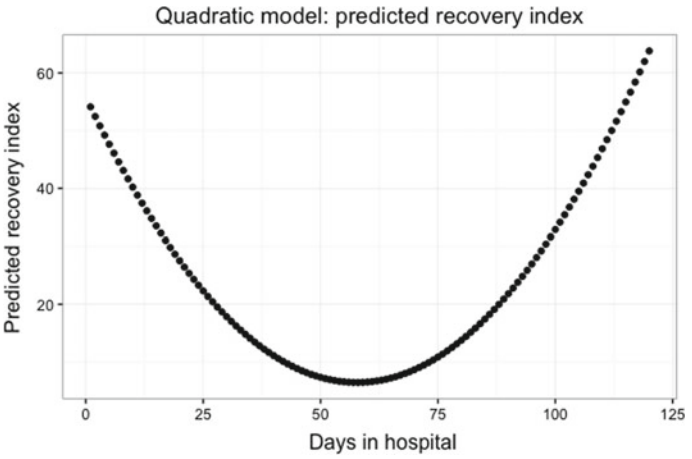


Fig. 8 Polynomial regression plot (quadratic polynomial)

4 Exponential Decay Modeling of Hospital Recovery Data

We may be able to build a model that both accurately fits the data and produces estimates that pass the common-sense test by using an exponential decay model. This modeling approach lets us model relationships that vary over time in a non-linear fashion—in this case, we want to accurately capture the strong correlation for lower ranges of T, but allow the magnitude of this relationship to decay as T increases, as the data seems to indicate.

Generating non-linear models in *R* is done using the non-linear least squares or NLS function, appropriately labeled *nls()*. This function automatically fits a wide range of non-linear models based on a functional form designated by the user. It is important to note that when fitting an NLS model in *R*, minimizing the sum of squares $\sum_{i=1}^{n}(y_i - a(exp(bx_i)))^2$ is done computationally rather than mathematically. That means that the choice of starting values for the optimization function is important—the estimates produced by the model may vary considerably based on the chosen starting values [4]! As such, it is wise to experiment when fitting these non-linear values to test how robust the resulting estimates are to the choice of starting values. We suggest using a ln-ln transformation of this data to begin with and then transforming back into the original xy space to obtain "good" estimates. The model, $ln(y) = ln(a) + bx$, yields $ln(y) = 4.037159 - 0.03797\ x$. This translates into the estimated model: $y = 56.66512e^{(-0.03797x)}$. Our starting values for (a, b) should be $(56.66512, -0.03797)$.

```
## Fit NLS model to the data
## Generate model
recovery_model3 <- nls(
  Y ~ a * (exp(b * T))
  , data = recovery_data
  , start = c(
    a=56.66512
    , b=-0.03797
  )
)
```

Fitting nonlinear regression model: *Y ~ a * (exp(b * T))*

Parameter Estimates

a	b
58.61	−0.03959

residual sum-of-squares: 1.951

The final model is $y = 58.61e^{-0.03959x}$. Overlaying the trend produced by the model on the plot of observed values, Fig. 8, we see that the NLS modeling approach fits the data very well (Fig. 9).

Once again, we can visually assess model fit by calculating and plotting the residuals. The figures below show the same residuals plotted along both days in the hospital T and recovery index Y.

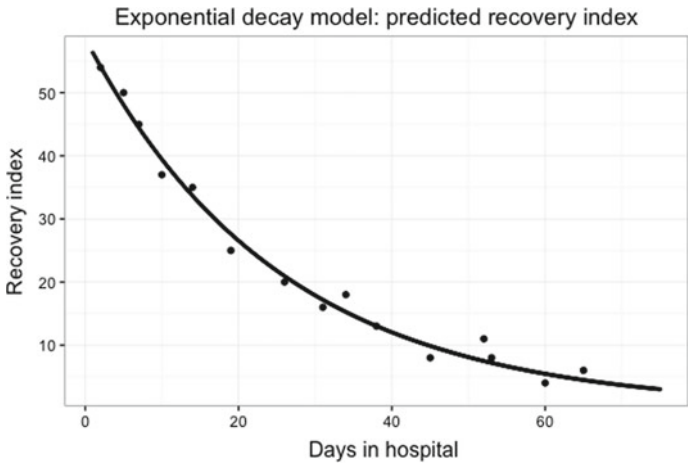


Fig. 9 Exponential regression model and data

```
## # A tibble: 15 × 5
##       T     Y predicted   residuals pct_relative_error
##   <int> <int>     <dbl>       <dbl>              <dbl>
## 1     2    54 54.143916 -0.14391597        -0.2665110
## 2     5    50 48.079006  1.92099420         3.8419884
## 3     7    45 44.418036  0.58196361         1.2932525
## 4    10    37 39.442567 -2.44256693        -6.6015322
## 5    14    35 33.664559  1.33544135         3.8155467
## 6    19    25 27.617389 -2.61738904       -10.4695562
## 7    26    20 20.931299 -0.93129928        -4.6564964
## 8    31    16 17.171407 -1.17140692        -7.3212932
## 9    34    18 15.247958  2.75204170        15.2891206
## 10   38    13 13.014259 -0.01425912        -0.1096856
## 11   45     8  9.863545 -1.86354475       -23.2943094
## 12   52    11  7.475609  3.52439082        32.0399165
## 13   53     8  7.185360  0.81464008        10.1830010
## 14   60     4  5.445805 -1.44580513       -36.1451283
## 15   65     6  4.467574  1.53242564        25.5404274
```

In both cases, we see that there is no easily distinguishable pattern in residuals. Finally, we apply the common-sense check by generating and plotting estimated recovery index values for a set of values of T from 1 to 120 (Fig. 10).

The predicted values generated by the exponential decay model make intuitive sense. As the number of days a patient spends in the hospital increases, the model predicts that their recovery index will decrease at a decreasing rate. This means that while the recovery index variable will continuously decrease, it will not take on negative values (as predicted by the linear model) or explosively large values (as predicted by the quadratic model). It appears that the exponential decay model not only fit the data best from a purely statistical point of view, but also generates values that pass the common-sense test to an observer or analyst shown in Fig. 11.

5 Sinusoidal Regression

Non-linear regression can also capture more complex functional forms, such as seasonal variation. In this section we analyze a set of monthly shipping data to understand how output relates to the current month. We begin as usual by inspecting the data as a table.

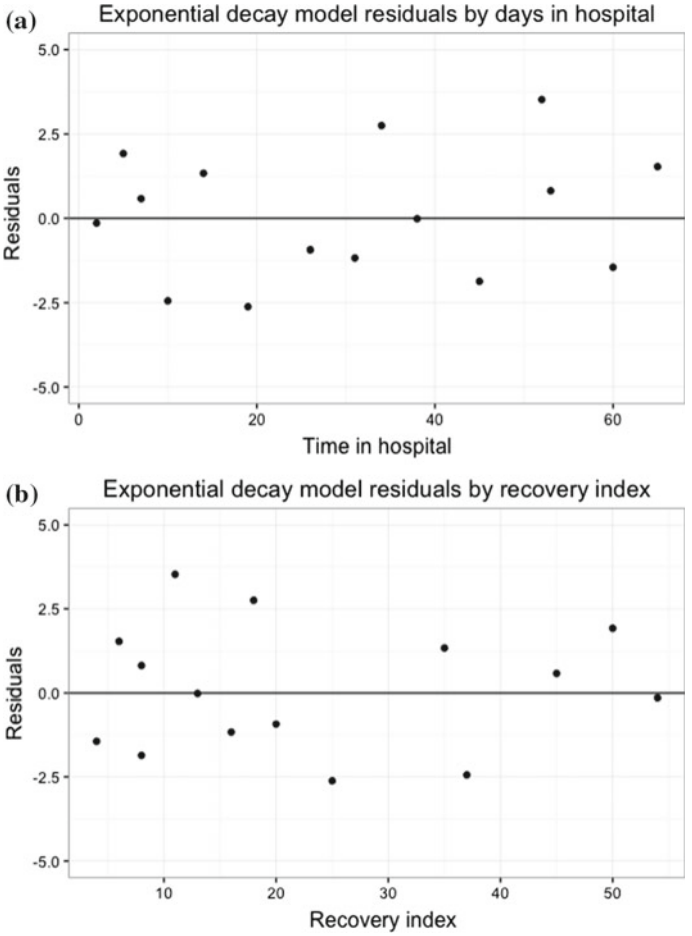


Fig. 10 **a** and **b** Residual plot as functions of the time and the model

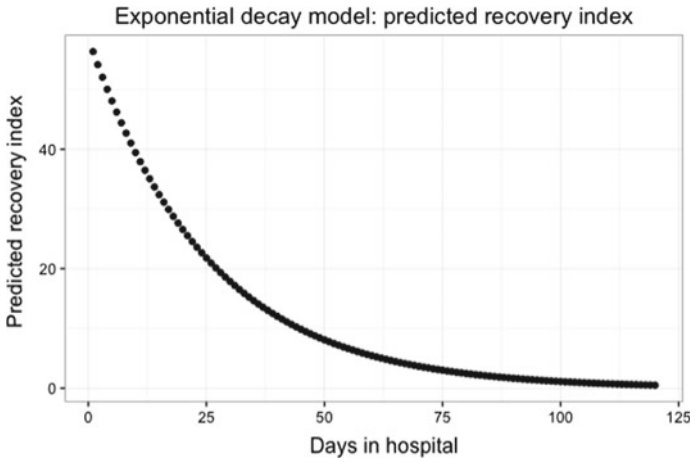


Fig. 11 Plot of exponential regression model

5.1 Introducing Shipping Data

```
## # A tibble: 20 × 2
##    Month UsageTons
##    <int>     <int>
## 1      1        20
## 2      2        15
## 3      3        10
## 4      4        18
## 5      5        28
## 6      6        18
## 7      7        13
## 8      8        21
## 9      9        28
## 10    10        22
## 11    11        19
## 12    12        25
## 13    13        32
## 14    14        26
## 15    15        21
## 16    16        29
## 17    17        35
## 18    18        28
## 19    19        22
## 20    20        32
```

```
## Calculate and print correlation matrix
print(cor(shipping_data))

##              Month UsageTons
## Month     1.0000000 0.6725644
## UsageTons 0.6725644 1.0000000
```

Once again, we can visualize the data in a scatter plot to assess whether this positive correlation is borne out by the overall trend (Fig. 12).

Visualizing the data, we see that there is a clear positive trend over time in shipping usage. However, examining the data in more detail suggests that a simple linear model may not be best-suited to capturing the variation in these data. One way to plot more complex patterns in data is through the use of a trend line using polynomial or non-parametric smoothing functions.

Plotting a trend line generated via a spline function shows that there seems to be an oscillating pattern with a steady increase over time in the shipping data (Fig. 13).

5.2 Linear Regression of Shipping Data

As a baseline for comparison, we begin by fitting a standard OLS regression model using the *lm()* function in *R*.

```
## Generate model
shipping_model1 <- lm(UsageTons ~ Month, data = shipping_data)
```

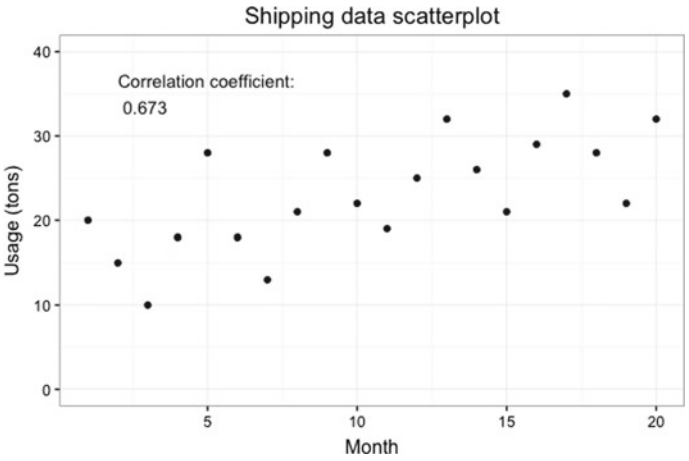


Fig. 12 Scatterplot of shipping data

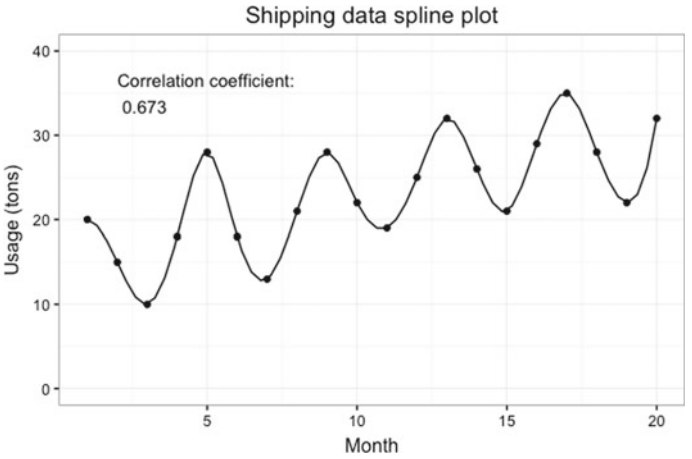


Fig. 13 Shipping data with data points connected show an oscillating trend

	Estimate	Std. error	t value	Pr(>|t|)
Month	0.7594	0.1969	3.856	0.001158
(Intercept)	15.13	2.359	6.411	4.907e−06

Fitting linear model: UsageTons ~ Month

Observations	Residual std. error	R^2	Adjusted R^2
20	5.079	0.4523	0.4219

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	1	383.5	383.5	14.87	0.001158
Residuals	18	464.3	25.79	NA	NA

While the linear model, $y = 15.13 + 0.7954\,x$, fits the data fairly well, the oscillation identified by the spline visualization suggests that we should apply a model that better fits the seasonal variation in the data.

5.3 Sinusoidal Regression of Shipping Data

R treats sinusoidal regression models as part of the larger family of nonlinear least-squares (NLS) regression models. This means that we can fit a sinusoidal model using the same *nls()* function and syntax as we applied earlier for the exponential decay model. The functional form for the sinusoidal model we use here can be written as:

$$Usage = a * sin(b * time + c) + d * time + e$$

This function can be expanded out trigonometrically as:

$$Usage = a * time + b * sin(c * time) + d * cos(c * time) + e$$

This equation can be passed to *nls()* and *R* will computationally assess best-fit values for the $a, b, c, d,$ and e terms. It is worth stressing again the importance of selecting good starting values for this process, especially for a model like this one with many parameters to be simultaneously estimated. Here, we set starting values based on pre-analysis of the data. It is also important to note that because the underlying algorithms used to optimize these functions differ between Excel and R, the two methods produce models with different parameters but nearly identical predictive qualities. The model can be specified in *R* as follows.

```
## Generate model
shipping_model2 <- nls(
  UsageTons ~ a * Month + b*sin(c*Month) + d*cos(c*Month) + e
  , data = shipping_data
  , start = c(
    a=5
    , b=10
    , c=1
    , d=1
    , e=10
    )
)
```

Fitting nonlinear regression model: UsageTons ~ a * Month + b * sin(c * Month) + d * cos(c * Month) + e

Parameter Estimates

a	*b*	*c*	*d*	*e*
0.848	*6.666*	*1.574*	*0.5521*	*14.19*

residual sum-of-squares: 1.206

The model found is:

$$Usage = 0.848 * time + 6.666 * sin(1.574 * time) + 0.5521 * cos(c(time)) + 14.19.$$

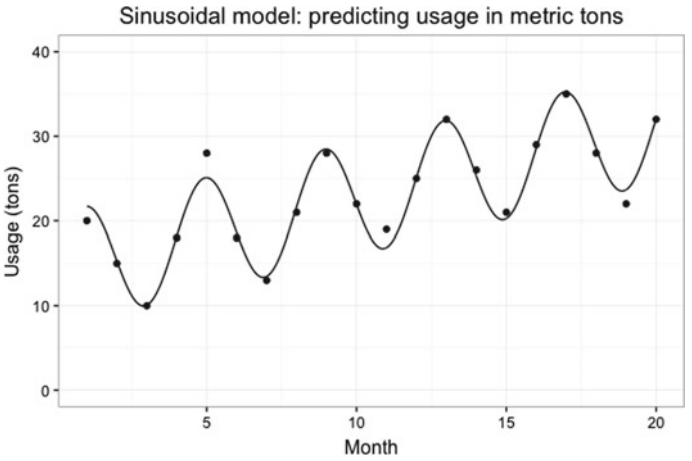


Fig. 14 Overlay of regression model and data

Plotting the trend line produced by the sinusoidal model shows that this modeling approach fits the data much better, accounting for both the short-term seasonal variation and the long-term increase in shipping usage (Fig. 14).

```
## # A tibble: 20 × 5
##    Month UsageTons predicted   residuals pct_relative_error
##    <int>    <int>     <dbl>       <dbl>              <dbl>
## 1      1       20  21.69933 -1.69932876         -8.4966438
## 2      2       15  15.29431 -0.29431044         -1.9620696
## 3      3       10  10.06872 -0.06872366         -0.6872366
## 4      4       18  18.20250 -0.20249983         -1.1249991
## 5      5       28  25.08458  2.91542390         10.4122282
## 6      6       18  18.61406 -0.61406056         -3.4114475
## 7      7       13  13.46748 -0.46747667         -3.5959744
## 8      8       21  21.66632 -0.66632268         -3.1729651
## 9      9       28  28.46904 -0.46904406         -1.6751573
## 10    10       22  21.93389  0.06611205          0.3005093
## 11    11       19  16.86701  2.13299137         11.2262704
## 12    12       25  25.13006 -0.13006404         -0.5202562
## 13    13       32  31.85273  0.14726650          0.4602078
## 14    14       26  25.25380  0.74619895          2.8699960
## 15    15       21  20.26732  0.73268136          3.4889588
## 16    16       29  28.59372  0.40628451          1.4009811
## 17    17       35  35.23565 -0.23564539         -0.6732725
## 18    18       28  28.57381 -0.57380826         -2.0493152
## 19    19       22  23.66841 -1.66840571         -7.5836623
## 20    20       32  32.05727 -0.05726862         -0.1789645
```

5.4 Introducing Afghanistan Casualty Data

See Fig. 15.

5.5 Sinusoidal Regression of Afghanistan Casualties

Visualizing data on casualties in Afghanistan between 2006 through 2008 shows an increasing trend overall, and significant seasonal oscillation. Once again, we want to fit a non-linear model that accounts for the oscillation present in the data. We use the same sinusoidal functional form

$$Casualties = a * sin(b * time + c) + d * time + e$$

which as before can be expressed as

$$Casualties = a * time + b * sin(c * time) + d * cos(c * time) + e$$

We fit the model using the *nls()* function once again:

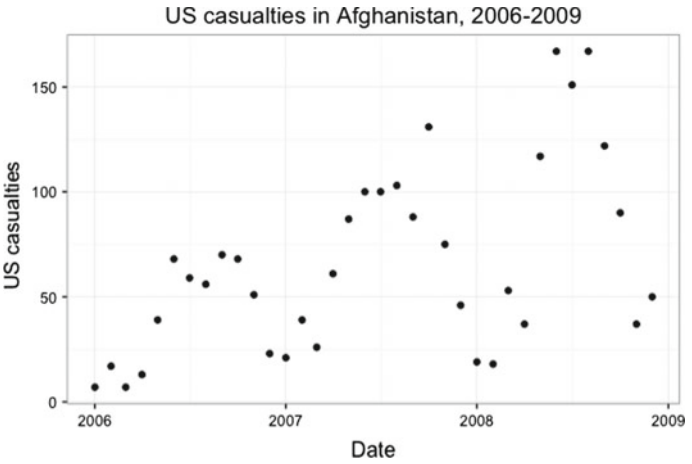


Fig. 15 Casualty data scatter plot

```
## Generate model
afghan_model <- nls(
  Casualties ~ a * DateIndex + b*sin(c*DateIndex) + d*cos(c*DateIndex) + e
  , data = afghan_data
  , start = c(
    a=1
    , b=10
    , c=1
    , d=10
    , e=1
  )
)
```

Fitting nonlinear regression model: Casualties ~ a * DateIndex + b * sin(c * DateIndex) + d * cos(c * DateIndex) + e

Parameter Estimates

a	*b*	*c*	*d*	*e*
1.85	*−42.92*	*0.547*	*−12.29*	*33.53*

residual sum-of-squares: 21.56

The model found is

$$Casualties = 1.85 * time \pm 42.92 * sin\,(0.547 * time) - 12.19 * cos\,(0.547 * time) + 33.53$$

Plotting the trend line identified by the sinusoidal model shows again that the sinusoidal modeling approach can account for both short-term oscillation and long-term increase. We can now estimate residuals and error metrics, and assess how well the model fits over the full range of the data (Figs. 16 and 17).

```
## # A tibble: 36 × 8
##    Year Month Casualties       Date DateIndex predicted    residuals
##   <int> <int>      <int>     <date>     <int>     <dbl>        <dbl>
## 1  2006     1          7 2006-01-01         1  2.559362   4.44063774
## 2  2006     2         17 2006-02-01         2 -6.539489  23.53948919
## 3  2006     3          7 2006-03-01         3 -2.862473   9.86247255
## 4  2006     4         13 2006-04-01         4 13.057033  -0.05703292
## 5  2006     5         39 2006-05-01         5 37.112404   1.88759617
## 6  2006     6         68 2006-06-01         6 62.822382   5.17761750
## 7  2006     7         59 2006-07-01         7 83.222777 -24.22277700
## 8  2006     8         56 2006-08-01         8 92.899116 -36.89911626
## 9  2006     9         70 2006-09-01         9 89.566985 -19.56698456
## 10 2006    10         68 2006-10-01        10 74.738779  -6.73877932
## # ... with 26 more rows, and 1 more variables: pct_relative_error <dbl>
```

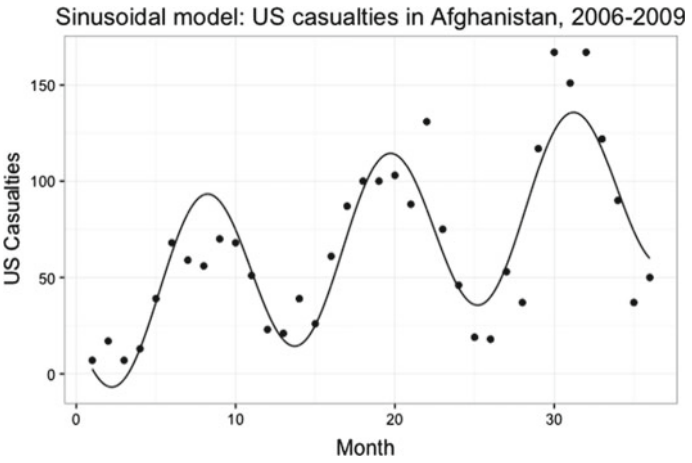


Fig. 16 Model of casualties

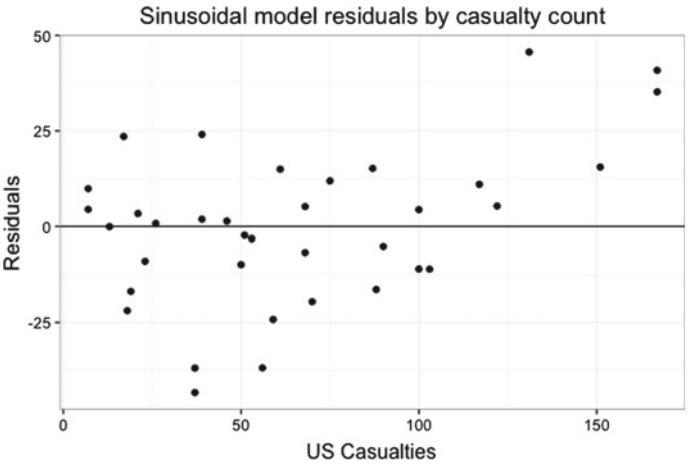


Fig. 17 Residual plot of casualty model

6 Logistic Regression

Often our dependent variable has special characteristics. Here is examine two such special cases: the dependent variable is binary {0,1} and the dependent variable is a count that follows a Poisson distribution.

6.1 Case Study: Dehumanization and the Outcome of Conflict with Logistics Regression

Dehumanization is not a new phenomenon in inter-human conflict. Man has arguably "dehumanized" his human adversaries to allow man to coerce, maim, or ultimately kill while avoiding the pain of conscience for committing the extreme, violent action. By taking away the human traits of his opponents, man has made his adversaries to be objects deserving of wrath and self-actualizing his justice of the action. Dehumanization still occurs today in both developed and underdeveloped societies within the inter-state system. This case analyzes the impact that dehumanization has, in its various manifested forms, on the outcome of a state's ability to win a conflict.

Data Specifics

To examine at dehumanization as a quantitative statistic, this case amalgamated data from a series of 25 conflicts and a previous study of civilian casualties from the respective conflicts. The conflict casualty data set derived from Erik Melander, Magnus Oberg and, Jonathan Hall's Uppsala Peace and Conflict research paper, "The 'New Wars' Debate Revisited: An Empirical Evaluation of the Atrociousness of 'New Wars'," is shown in Fig. 18.

As stated earlier, the above conflicts represent the high- and low-intensity spectrum of conflict, and include both inter- and intra-state conflicts. Thus the data is a fair representation of conflict in general. However, the above data table was used in support of a study that focused on the casualty output of conflict and not on the interrelation of civilian casualties that we define as an indicator of dehumanization to the outcome of the conflict for the state. Typically, there is no unambiguous victor or vanquished in conflict, but to allow us to analyze the relationship of civilian casualty ratios and the outcome of the conflict it was necessary to utilize a definitive binary assessment of each of the above conflicts' winners and losers. To this end we utilized an additional data set that codified conflicts in terms of two sides with the determination of which side "won" each respective conflict. The implications of this case study vary broadly, but we were singularly focused on civilian deaths in conflict as an indicator of dehumanization's occurrence, and subsequently dehumanization's effect on the state's ability to win the conflict.

By taking a ratio of the civilian casualties in relationship to the total casualties we were able to determine what percentages of casualties in each conflict were civilian, shown in Fig. 18. This provided us a quantifiable independent variable to analyze. Additionally, we made the inference that the conflicts with higher civilian casualty percentages likely incurred a higher amount of "value targeting," a previously discussed symptom of dehumanization. By using the civilian casualty percentage independent variable and comparing it to the assessed binary outcome of either a win or loss as the dependent variable, we were able to synthesize the data into a binary logistical regression model to assess the significance of the civilian casualty percentages on the outcome of the state's (*Side A*) ability win the conflict. For more information see Kreutz [9]. Data is provided in Fig. 19.

Country	*Year*	*Civilian*	*Military*	*Total*
India	1946-48	800,000	0	800,000
Columbia	1949-62	200,000	100,000	300,000
China	1950-51	1,000,000	*	1,000,000
Korea	1950-53	1,000,000	1,889,000	2,889,999
Algeria	1954-62	82,000	18,000	100,000
Tibet	1956-59	60,000	40,000	100,000
Rwanda	1956-65	102,000	3,000	105,000
Iraq	1961-70	100,000	5,000	105,000
Sudan	1963-72	250,000	250,000	500,000
Indonesia	1965-66	500,000	*	500,000
Vietnam	1965-75	1,000,000	1,058,000	2,058,000
Guatemala	1966-87	100,000	38,000	138,000
Nigeria	1967-70	1,000,000	1,000,000	2,000,000
Egypt	1967-70	50,000	25,000	75,000
Bangladesh	1971-71	1,000,000	500,000	1,500,000
Uganda	1971-78	300,000	0	300,000
Burundi	1972-72	80,000	20,000	100,000
Ethiopia	1974-87	500,000	46,000	546,000
Lebanon	1975-76	76,000	25,000	100,000
Cambodia	1975-78	1,500,000	500,000	2,000,000
Angola	1975-87	200,000	13,000	213,000
Afghanistan	1978-87	50,000	50,000	100,000
El Salvador	1979-87	50,000	15,000	65,000
Uganda	1981-87	100,000	2,000	102,000
Mozambique	1981-87	350,000	51,000	401,000

Source: Adapted from World Military and Social Expenditures 1987-88 (Sivard, 1987). * denotes missing values.

Fig. 18 Civilian and military casualties resultant from high- and low-intensity conflicts (*Source* [10])

6.2 A Binary Logistical Regression Analysis of Dehumanization

Binary logistical regression analysis is an ideal method to analyze the interrelation of dehumanization's effects (shown through higher percentages of civilian casualties) on the outcome of conflict (shown to be a win "1" or a loss "0"). Binary logistical regression model statistics will allow us to explain whether or not the civilian

Conflict's Country Location	Side A	Side A Win (1) or Loss (0)	Side B	Side B Win (1) or Loss (0)	Year	Civilian Casualties	Military Casualties	Total Casualties	Civilian Deaths Percentage
India	India	1	CPI	0	1946-48	800,000	0	800,000	1.0000
Columbia	Columbia	1	Military Junta	0	1949-62	200,000	100,000	300,000	0.6667
China	China	1	Taiwan	0	1950-51	1,000,000	*	1,000,000	1.0000
Korea	North Korea	0	South Korea	1	1950-53	1,000,000	1,889,000	2,889,999	0.3460
Algeria	France	0	FLN	1	1954-62	82,000	18,000	100,000	0.8200
Tibet	China	1	Tibet	0	1956-59	60,000	40,000	100,000	0.6000
Rwanda	Tutsi	0	Hutu	1	1956-65	102,000	3,000	105,000	0.9714
Iraq	Iraq	1	KDP	0	1961-70	100,000	5,000	105,000	0.9524
Sudan	Sudan	1	Anya Nya	0	1963-72	250,000	250,000	500,000	0.5000
Indonesia	Indonesia	1	OPM	0	1965-66	500,000	*	500,000	1.0000
Vietnam	North Vietnam	1	South Vietnam	0	1965-75	1,000,000	1,058,000	2,058,000	0.4859
Guatemala	Guatamala	1	FAR	0	1966-87	100,000	38,000	138,000	0.7246
Nigeria	Nigeria	1	Republic of Biafra	0	1967-70	1,000,000	1,000,000	2,000,000	0.5000
Egypt	Egypt	0	Israel	1	1967-70	50,000	25,000	75,000	0.6667
Bangladesh	Bangladesh	1	JSS/SB	0	1971-71	1,000,000	500,000	1,500,000	0.6667
Uganda	Uganda	1	Military Faction	0	1971-78	300,000	0	300,000	1.0000
Burundi	Burundi	1	Military Faction	0	1972-72	80,000	20,000	100,000	0.8000
Ethiopia	Ethiopia	1	OLF	0	1974-87	500,000	46,000	546,000	0.9158
Lebanon	Lebanon	1	LNM	0	1975-76	76,000	25,000	100,000	0.7600
Cambodia	Cambodia	0	Khmer Rouge	1	1975-78	1,500,000	500,000	2,000,000	0.7500
Angola	Angola	1	FNLA	0	1975-87	200,000	13,000	213,000	0.9390
Afghanistan	Afghanistan	1	USSR	0	1978-87	50,000	50,000	100,000	0.5000
El Salvador	El Salvador	1	FMLN	0	1979-87	50,000	15,000	65,000	0.7692
Mozambique	Mozambique	1	Renamo	0	1981-87	350,000	51,000	401,000	0.8728
Uganda	Uganda	1	Kikosi Maalum et al.	0	1981-87	100,000	2,000	102,000	0.9804

Fig. 19 Conflict outcomes and civilian casualty percentages data set (*Source* [9])

casualties percentage (independent variable) has a significance level on the outcome. Using the data table from figure two, we assessed the civilian casualty percentages to be the independent variable "X" and *Side A*'s win/loss outcome from the conflict to be the dependent variable "Y." From this data we were able to develop a binary logistical regression model. Using statistical analysis software package, we derived the logistics regression statistics from the model, shown from Minitab©, Fig. 19.

Conflict outcomes differ from the data we have examined so far in that the measure of state victory only has two values, 1 and 0. This type of data be modeled using a binomial logistic (or sometimes "logit") regression. Logistic regression estimates an underlying continuous variable usually referred to as Y^* that is then transformed into an estimate bounded below by 0 and above by 1. This means the logistic modeling approach is extremely useful for estimating binary (1/0) outcomes, as the estimated values can be easily translated into either point estimates or log-probabilities of observing a 1 versus a 0:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

The logistic model in *R* is treated as one case of a broader range of generalized linear models (GLM), and can be accessed via the conveniently named *glm()* function. Note that because *glm()* implements a wide range of generalized linear models based on the inputs provided, it is necessary for the user to specify both the family of model (binomial) and the link function (logit).

```
## Generate model
war_model <- glm(
  side_a ~ cd_pct
  , data = war_data
  , family = binomial(link = 'logit')
  )
```

Fitting generalized (binomial/logit) linear model: side_a ~ cd_pct

	Estimate	Std. error	z value	Pr(>|z|)
cd_pct	1.85	2.556	0.7237	0.4692
(Intercept)	0.004716	1.925	0.00245	0.998

Logistic regression shows that there is a positive correlation between civilian casualties and state victory, but that this relationship is not statistically significant at the $p < 0.05$ level. This means we cannot reject the null hypothesis H_0 that no relationship exists between the input and output variables.

6.3 Introducing International Alliance Data

We now turn to a larger data set, measuring alliance connections between politically relevant states (powerful states and those that share a border with one another) in the international system in the year 2000. Scholars are often interested in assessing the factors that predict whether two states will form a military alliance, as these are salient and lasting forms of cooperation that signal trust (or at least, a lack of overt enmity) between governments.

Coupled with data on whether or not an alliance exists, we also have data on the level to which each pair of states share membership in major intergovernmental organizations (IGOs). These IGOs include major international entities such as the United Nations, the World Trade Organization, and the International Atomic Energy Agency, as well as regional or policy-based organizations such as the Association of Southeast Asian Nations (ASEAN) or the Organization of Petroleum Exporting Countries (OPEC).

The data used for this analysis is presented in the table below. The first two columns identify the ISO-3000 code identifying each country. Alliances are recorded as being present (1) or absent (0), and the overlap of IGO membership is recorded as a count value bounded below by zero.

```
## # A tibble: 1,586 × 4
##    statea stateb alliance_present igo_overlap
##    <chr>  <chr>           <int>      <int>
## 1    AZE    ARM               1         33
## 2    BFA    BEN               1         67
## 3    BOL    ARG               1         63
## 4    BRA    ARG               1         73
## 5    BRA    BOL               1         64
## 6    CHE    AUT               0         74
## 7    CHL    ARG               1         73
## 8    CHL    BOL               1         63
## 9    CHN    AFG               0         27
## 10   CHN    AGO               0         29
## # ... with 1,576 more rows
```

6.4 Logistic Regression of Alliance Data

States which share membership in many of the same IGOs are likely to have similar policy preferences, regional concerns, and economic status that lead to their choosing to join these organizations. If we believe that similarity breeds familiarity and lowers barriers to cooperation (similar to the 'birds of a feather' argument), then we can generate testable expectations about how shared IGO membership relates to the probability of forming an alliance between states. Specifically, we hypothesize that as shared IGO membership between a pair of states increases, the probability that these states also share a military alliance will increase as well.

We can test this hypothesis by fitting another logistic model in *R* using the *glm()* function.

```
alliance_model <- glm(
  alliance_present ~ igo_overlap
  , data = alliance_data
  , family = binomial(link = 'logit')
  )
```

	Estimate	Std. error	z value	Pr(>|z|)
igo_overlap	0.08358	0.005461	15.3	7.2e−53
(Intercept)	−5.121	0.2617	−19.57	2.937e−85

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1497 on 1585 degrees of freedom
Residual deviance:	1156 on 1584 degrees of freedom

The results of the logistic regression suggest that there is a positive relationship between the number of IGO memberships a pair of states share and the likelihood that they also share an alliance. This relationship is significant at the $p < 0.01$ level, meaning that we can reject the null hypothesis H_0 with a high level of confidence.

Remember that logistic regression models can produce estimated probabilities of observing a 1 versus a 0 based on a given set of input values. This is a useful way of visualizing how well a model fits the observed data. Here, we produce a set of predicted probabilities (bounded between 0 and 1) that an alliance will be present between each pair of states based on their IGO membership overlap, and overlay this trend line on the scatter plot of 0 and 1 values present in the data. The plot is shown in Fig. 20.

Visualizing the predicted probability estimates shows that the model does a moderately good job of separating out 0's and 1's based on the inputs used. IGO membership is certainly not the only factor that may explain how states form alliances with one another, but it provides a useful starting point for modeling.

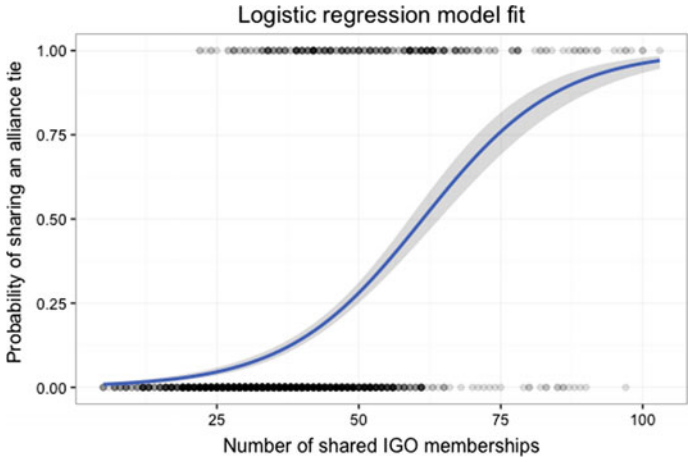


Fig. 20 Logistic model for IGO membership

7 Poisson Regression

7.1 Introducing SIGACTS Data

As discussed earlier in the chapter, the regional SIGACTS data recorded in the Philippines are count data, meaning they take only integer values and are bounded below by zero. Visualizing count data in a histogram is a useful way of assessing how the data are distributed.

Visualizing the data in a histogram we observe that they appear be Poisson-distributed, which is common in count data. We also recommend applying a goodness of fit test to prove the data is Poisson. The histogram in Fig. 21 appears to look Poisson. The goodness of fit test does confirm a Poisson distribution.

7.2 Poisson Regression of SIGACTS Data

Poisson regression in *R* is also treated as a special case of GLMs, similar to the logistic regression covered in the previous section. As such, it can be implemented using the same *glm()* function, but now specifying the model family as 'Poisson', which tells R to implement a Poisson model. The model we use here can be specified as

$$Y = e^{\beta_0+\beta_1 GGI+\beta_2 Literacy+\beta_3 Poverty}$$

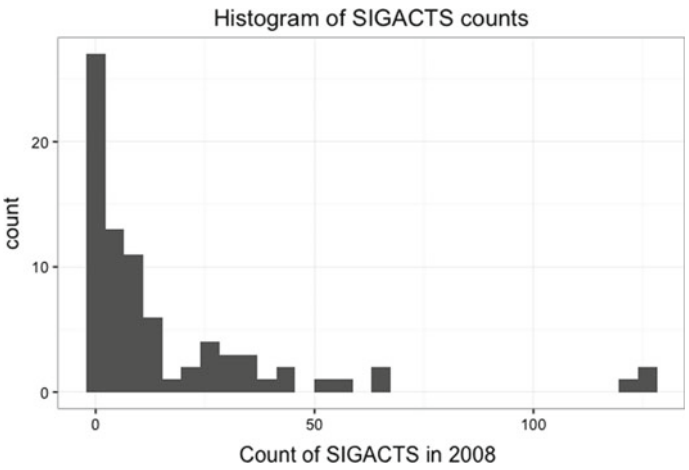


Fig. 21 Histogram of SIGACTS in 2008

```
## Generate model
sigacts_model <- glm(
  sigacts_2008 ~ ggi_2008 + literacy + poverty
  , data = sigacts_data
  , family = poisson
)
```

	Estimate	Std. error	z value	Pr(>|z|)
ggi_2008	−0.0136	0.001475	−9.22	2.973e−20
literacy	−0.02098	0.005091	−4.12	3.79e−05
poverty	0.02297	0.002214	10.37	3.265e−25
(Intercept)	5.288	0.4665	11.34	8.755e−30

(Dispersion parameter for Poisson family taken to be 1)

Null deviance:	2358 on 79 degrees of freedom
Residual deviance:	1852 on 76 degrees of freedom

The model is $SIGACTS = e^{(5.288+0.02297\ Poverty-0.02098\ Literacy-0.0136\ ggi)}$.

Note that Poisson models generate log-odds estimates. This means that we can readily convert coefficient estimates to odds ratios, indicating the impact that a one-unit change in a given input variable will have on the estimated number of events. When interpreting odds ratios, remember that an odds ratio above 1.0 indicates that increasing the input variable increase the estimated event count, while odds ratios lower than 1.0 indicate that increasing the input variable will lower the estimated event count.

- $\exp(-0.0136) = 0.986$. This means that increasing the value of government satisfaction by one unit will lower the expected level of violence by about 1.4%.
- $\exp(-0.02098) = 0.979$. This means that increasing the value of literacy by one unit will lower the expected level of violence by about 2.1%.
- $\exp(0.02297) = 1.023$. This means that increasing the value of poverty by one unit will increase the expected level of violence by about 1.02%.

These relationships are all in the direction we would intuitively expect: higher literacy and greater satisfaction with the government should certainly be associated with lower levels of anti-government violence, while greater poverty may drive discontent and disorder, including violent acts. All three estimated relationships are statistically significant at the $p < 0.01$ level, meaning that although the magnitude of change is not large, we can safely reject the null hypothesis that no relationship exists.

8 Conclusions and Summary

We showed some of the common misconceptions by decision makers concerning correlation and regression. Our purpose of this presentation is to help prepare more competent and confident problem solvers for the 21st century. Data can be found using part of a sine curve where the correlation is quite poor, close to zero but the decision maker can describe the pattern. Decision makers see the relationship in the data as periodic or oscillating. Examples such as these should dispel the idea that correlation of almost zero implies no relationship. Decision makers need to see and believe concepts concerning correlation, linear relationships, and non-linear (or no) relationship.

We recommended the following summary steps.

Step 1. Insure you understand the problem and what answers are required.
Step 2. Get the data that is available. Identify the dependent and independent variables.
Step 3. Plot the dependent versus an independent variable and note trends.
Step 4. If the dependent variable is binary {0,1} then use binary logistics regression. If the dependent variables are counts that follow a Poisson distribution, then use Poisson regression. Otherwise, try linear, multiple, or nonlinear regression as needed.
Step 5. Insure your model produces results that are acceptable.

9 Using R

Before we start: setting up the workspace
Before working in *R*, it is necessary to set up the "workspace": the virtual environment in which you can load, manipulate, and analyze data. The code below cleans the workspace, erasing any previous objects or functions; sets the working directory, from which we'll load the data to analyze, and also loads a set of "packages" of useful functions that make data cleaning and analysis easier and faster.

Using the *cor()* command in *R* on the data table:

```
############################
##
## Setting up workspace
##
############################

## Clear previous workspace, if any
rm(list=ls())

## Set working directory
os_detect <- Sys.info()['sysname']
if (os_detect == 'Darwin'){
  setwd('/Users/localadmin/Dropbox/Research/StatsChapter')
}

## Load packages for analysis
pacman::p_load(
  data.table, tidyverse, ggplot2, stargazer, easynls, pscl, pander
)

############################
##
## Reading in the data sets used in this chapter
##
############################

## Read in spring data
spring_data <- read_csv('./Data/01_correlation.csv')

## Parsed with column specification:
## cols(
##   x = col_integer(),
##   y = col_double()
## )

sigacts_data <- read_csv('./Data/06_poisson_sigacts.csv')
```

```
## Parsed with column specification:
## cols(
##   sigacts_2008 = col_integer(),
##   ggi_2008 = col_double(),
##   literacy = col_double(),
##   poverty = col_double()
## )

recovery_data <- read_csv('./Data/02_exponential_decay.csv')

## Parsed with column specification:
## cols(
##   T = col_integer(),
##   Y = col_integer()
## )

shipping_data <- read_csv('./Data/03_sine_regression_shipping.csv')

## Parsed with column specification:
## cols(
##   Month = col_integer(),
##   UsageTons = col_integer()
## )

afghan_data <- read_csv('./Data/04_sine_regression_casualties.csv')

## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Month = col_integer(),
##   Casualties = col_integer()
## )

war_data <- read_csv('./Data/05_bin_logit_conflict.csv')

## Parsed with column specification:
## cols(
##   side_a = col_integer(),
##   cd_pct = col_double()
## )

sigacts_data <- read_csv('./Data/06_poisson_sigacts.csv')

## Parsed with column specification:
## cols(
##   sigacts_2008 = col_integer(),
##   ggi_2008 = col_double(),
##   literacy = col_double(),
```

```
##   poverty = col_double()
## )

alliance_data <- read_csv('./Data/07_bin_logit_alliance.csv')

## Parsed with column specification:
## cols(
##   statea = col_character(),
##   stateb = col_character(),
##   alliance_present = col_integer(),
##   igo_overlap = col_integer()
## )

## Format and subset casualties data
afghan_data <- mutate(
  afghan_data
  , Date = as.Date(paste0(Year, '-', Month, '-', '01'), format = '%Y-%m-%d')
  ) %>% filter(
    Date >= as.Date('2006-01-01')
    , Date <= as.Date('2008-12-01')
  ) %>% mutate(
    DateIndex = 1:36
  )

## Print data as a tibble
print(spring_data)

## # A tibble: 11 × 2
##        x     y
##    <int> <dbl>
## 1     50 0.1000
## 2    100 0.1875
## 3    150 0.2750
## 4    200 0.3250
## 5    250 0.4375
## 6    300 0.4875
## 7    350 0.5675
## 8    400 0.6500
## 9    450 0.7250
## 10   500 0.8000
## 11   550 0.8750
```

9.1 Correlation in R

Using the *cor()* command in *R* on the data table:

```
## Calculate and print correlation matrix
print(cor(spring_data))
```

```
##          x        y
## x 1.0000000 0.9992718
## y 0.9992718 1.0000000
```

9.2 Plotting in R

```
## Generate a plot visualizing the data
spring_cor_plot <- ggplot(
  aes(x = x, y = y)
  , data = spring_data) +
  geom_point() +
  annotate(
    'text'
    , x = 100
    , y = 0.75
    , label = 'Correlation coefficient:\n 0.999272'
    , hjust = 0) +
  ggtitle('Spring data scatterplot') +
  theme_bw()

## Print the plot to console
plot(spring_cor_plot)
```

9.3 Fitting an Ordinary Least-Squares (OLS) Model with Form y y + ϵ to the Spring Data in R

```
## Fit OLS model to the data
spring_model <- lm(
  y ~ x
  , data = spring_data
  )
```

9.4 Correlation Matrix in R

```
## Calculate and print correlation matrix
print(cor(recovery_data))
```

```
##           T         Y
## T  1.0000000 -0.9410528
## Y -0.9410528  1.0000000
```

9.5 Quadratic Regression of Hospital Recovery Data

```
## Generate model
recovery_model2 <- lm(Y ~ T + I(T^2), data = recovery_data)
```

```
## # A tibble: 15 × 6
##        T     Y index predicted  residuals pct_relative_error
##    <int> <int> <int>     <dbl>      <dbl>              <dbl>
## 1      2    54     1 52.460836  1.5391644          2.8503045
## 2      5    50     2 47.640993  2.3590072          4.7180144
## 3      7    45     3 44.575834  0.4241663          0.9425917
## 4     10    37     4 40.200199 -3.2001992         -8.6491871
## 5     14    35     5 34.780614  0.2193857          0.6268164
## 6     19    25     6 28.672445 -3.6724455        -14.6897820
## 7     26    20     7 21.364792 -1.3647924         -6.8239618
## 8     31    16     8 17.033457 -1.0334567         -6.4591042
## 9     34    18     9 14.790022  3.2099781         17.8332119
## 10    38    13    10 12.213370  0.7866302          6.0510012
## 11    45     8    11  8.844363 -0.8443634        -10.5545422
## 12    52    11    12  6.926437  4.0735627         37.0323886
## 13    53     8    13  6.770903  1.2290967         15.3637082
## 14    60     4    14  6.511355 -2.5113548        -62.7838691
## 15    65     6    15  7.214379 -1.2143795        -20.2396576
```

9.6 Prediction

```
## Create a set of hypothetical patient observations with days in the hospital from 1
to 120
patient_days = tibble(T = 1:120)
```

```
## Feed the new data to the model to generate predicted recovery index values
predicted_values = predict(
  recovery_model2
  , newdata = patient_days
  )
```

9.7 Nonlinear Regression

9.7.1 Exponential Decay Modeling of Hospital Recovery Data

```
## Fit NLS model to the data
## Generate model
recovery_model3 <- nls(
  Y ~ a * (exp(b * T))
  , data = recovery_data
  , start = c(
    a=1
    , b=0.05
  )
  , trace = T
)
```

9.7.2 Sinusoidal Regression

The functional form for the sinusoidal model we use here can be written as:

$$Usage = a * sin(b * time + c) + d * time + e$$

This function can be expanded out trigonometrically as:

$$Usage = a * time + b * sin(c * time) + d * cos(c(time)) + e$$

```
## Generate model
shipping_model2 <- nls(
  UsageTons ~ a * Month + b*sin(c*Month) + d*cos(c*Month) + e
  , data = shipping_data
  , start = c(
    a=5
    , b=10
    , c=1
    , d=1
    , e=10
    )
  , trace = T
)
```

9.7.3 Sinusoidal Regression of Afghanistan Casualties

Visualizing data on casualties in Afghanistan between 2006 and 2008 shows an increasing trend overall, and significant seasonal oscillation. Once again, we want to fit a non-linear model that accounts for the oscillation present in the data. We use the same sinusoidal functional form

$$Casualties = a * sin(b * time + c) + d * time + e$$

which as before can be expressed as

$$Casualties = a * time + b * sin(c * time) + d * cos(c * time) + e$$

The logistic model in *R* is treated as one case of a broader range of generalized linear models (GLM), and can be accessed via the conveniently named *glm()* function. Note that because *glm()* implements a wide range of generalized linear models based on the inputs provided, it is necessary for the user to specify both the family of model (binomial) and the link function (logit).

```
## Generate model
war_model <- glm(
  side_a ~ cd_pct
  , data = war_data
  , family = binomial(link = 'logit')
  )
```

9.7.4 Poisson Regression

Visualizing count data in a histogram is a useful way of assessing how the data are distributed.

9.8 Histogram Plot

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

Poisson regression in *R* is also treated as a special case of GLMs, similar to the logistic regression covered in the previous section. As such, it can be implemented using the same *glm()* function, but now specifying the model family as 'Poisson', which tells R to implement a Poisson model. The model we use here can be specified as

$$Y = e^{\beta_0 + \beta_1 GGI + \beta_2 Literacy + \beta_3 Poverty}$$

```
## Generate model
sigacts_model <- glm(
  sigacts_2008 ~ ggi_2008 + literacy + poverty
  , data = sigacts_data
  , family = poisson
)
```

References and Suggested Reading

1. Affi, A., & Azen, S. (1979). *Statistical analysis* (2nd ed., pp. 143–144). London, UK: Academic Press.
2. Devore, J. (2012). *Probability and statistics for engineering and the sciences* (8th ed., pp. 211–217). Belmont, CA: Cengage Publisher.
3. Fox, W. P., & Fowler, C. (1996). Understanding covariance and correlation. *PRIMUS, VI*(3), 235–244.
4. Fox, W. (2012). *Mathematical modeling with maple*. Boston, MA: Cengage Publishers.
5. Fox, W. P. (2011, October–December). Using the EXCEL solver for nonlinear regression. *Computers in Education Journal (COED), 2*(4), 77–86.
6. Fox, W. P. (2012). Issues and importance of "good" starting points for nonlinear regression for mathematical modeling with maple: Basic model fitting to make predictions with oscillating data. *Journal of Computers in Mathematics and Science Teaching., 31*(1), 1–16.
7. Giordano, F., Fox, W., & Horton, S. (2013). *A first course in mathematical modeling* (5th ed.). Boston, MA: Cengage Publishers.
8. Johnson, I. (2012). An introductory handbook on probability, statistics, and excel. Retrieved July 11, 2012, from http://records.viu.ca/~johnstoi/maybe/maybe4.htm.
9. Kreutz, J. (2010). How and When Armed Conflicts End: Introducing the UCDP Conflict Termination Dataset. *Journal of Peace Research 47*(2), 243–250.
10. Melander, E., Oberg, M. & Hall, J. (2006). *The 'New Wars' debate revisited: An empirical evaluation of the atrociousness of 'New Wars'*. Uppsala peace research papers no. 9, department of peace and conflict research. Sweden: Uppsala University. http://www.musik.uu.se/digitalAssets/18/18585_UPRP_No_9.pdf. Accessed 12 Sept 2012.
11. Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed., pp. 531–547). Chicago, Il: Irwin Press.

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1497 on 1585 degrees of freedom
Residual deviance:	1156 on 1584 degrees of freedom

The results of the logistic regression suggest that there is a positive relationship between the number of IGO memberships a pair of states share and the likelihood that they also share an alliance. This relationship is significant at the $p < 0.01$ level, meaning that we can reject the null hypothesis H_0 with a high level of confidence.

Remember that logistic regression models can produce estimated probabilities of observing a 1 versus a 0 based on a given set of input values. This is a useful way of visualizing how well a model fits the observed data. Here, we produce a set of predicted probabilities (bounded between 0 and 1) that an alliance will be present between each pair of states based on their IGO membership overlap, and overlay this trend line on the scatter plot of 0 and 1 values present in the data. The plot is shown in Fig. 20.

Visualizing the predicted probability estimates shows that the model does a moderately good job of separating out 0's and 1's based on the inputs used. IGO membership is certainly not the only factor that may explain how states form alliances with one another, but it provides a useful starting point for modeling.

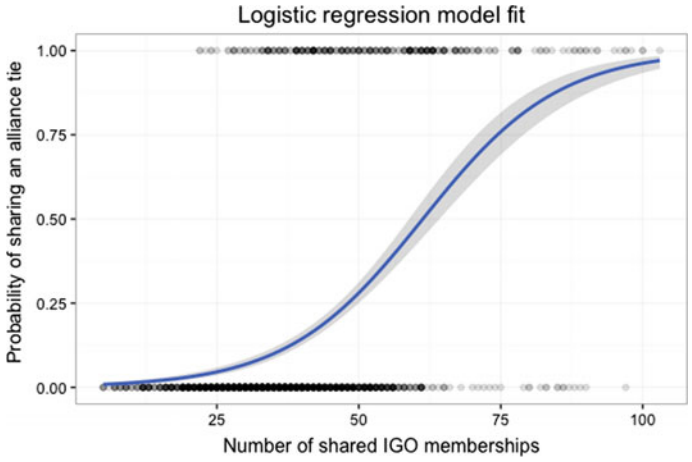


Fig. 20 Logistic model for IGO membership

7 Poisson Regression

7.1 Introducing SIGACTS Data

As discussed earlier in the chapter, the regional SIGACTS data recorded in the Philippines are count data, meaning they take only integer values and are bounded below by zero. Visualizing count data in a histogram is a useful way of assessing how the data are distributed.

Visualizing the data in a histogram we observe that they appear be Poisson-distributed, which is common in count data. We also recommend applying a goodness of fit test to prove the data is Poisson. The histogram in Fig. 21 appears to look Poisson. The goodness of fit test does confirm a Poisson distribution.

7.2 Poisson Regression of SIGACTS Data

Poisson regression in *R* is also treated as a special case of GLMs, similar to the logistic regression covered in the previous section. As such, it can be implemented using the same *glm()* function, but now specifying the model family as 'Poisson', which tells R to implement a Poisson model. The model we use here can be specified as

$$Y = e^{\beta_0+\beta_1 GGI+\beta_2 Literacy+\beta_3 Poverty}$$

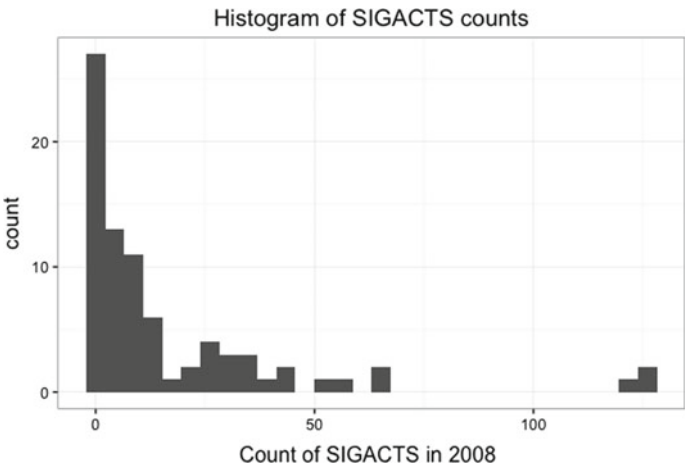


Fig. 21 Histogram of SIGACTS in 2008

```
## Generate model
sigacts_model <- glm(
  sigacts_2008 ~ ggi_2008 + literacy + poverty
  , data = sigacts_data
  , family = poisson
)
```

	Estimate	Std. error	z value	Pr(>|z|)
ggi_2008	−0.0136	0.001475	−9.22	2.973e−20
literacy	−0.02098	0.005091	−4.12	3.79e−05
poverty	0.02297	0.002214	10.37	3.265e−25
(Intercept)	5.288	0.4665	11.34	8.755e−30

(Dispersion parameter for Poisson family taken to be 1)

Null deviance:	2358 on 79 degrees of freedom
Residual deviance:	1852 on 76 degrees of freedom

The model is $SIGACTS = e^{(5.288+0.02297\ Poverty-0.02098\ Literacy-0.0136\ ggi)}$.

Note that Poisson models generate log-odds estimates. This means that we can readily convert coefficient estimates to odds ratios, indicating the impact that a one-unit change in a given input variable will have on the estimated number of events. When interpreting odds ratios, remember that an odds ratio above 1.0 indicates that increasing the input variable increase the estimated event count, while odds ratios lower than 1.0 indicate that increasing the input variable will lower the estimated event count.

- $\exp(-0.0136) = 0.986$. This means that increasing the value of government satisfaction by one unit will lower the expected level of violence by about 1.4%.
- $\exp(-0.02098) = 0.979$. This means that increasing the value of literacy by one unit will lower the expected level of violence by about 2.1%.
- $\exp(0.02297) = 1.023$. This means that increasing the value of poverty by one unit will increase the expected level of violence by about 1.02%.

These relationships are all in the direction we would intuitively expect: higher literacy and greater satisfaction with the government should certainly be associated with lower levels of anti-government violence, while greater poverty may drive discontent and disorder, including violent acts. All three estimated relationships are statistically significant at the $p < 0.01$ level, meaning that although the magnitude of change is not large, we can safely reject the null hypothesis that no relationship exists.

8 Conclusions and Summary

We showed some of the common misconceptions by decision makers concerning correlation and regression. Our purpose of this presentation is to help prepare more competent and confident problem solvers for the 21st century. Data can be found using part of a sine curve where the correlation is quite poor, close to zero but the decision maker can describe the pattern. Decision makers see the relationship in the data as periodic or oscillating. Examples such as these should dispel the idea that correlation of almost zero implies no relationship. Decision makers need to see and believe concepts concerning correlation, linear relationships, and non-linear (or no) relationship.

We recommended the following summary steps.

Step 1. Insure you understand the problem and what answers are required.
Step 2. Get the data that is available. Identify the dependent and independent variables.
Step 3. Plot the dependent versus an independent variable and note trends.
Step 4. If the dependent variable is binary {0,1} then use binary logistics regression. If the dependent variables are counts that follow a Poisson distribution, then use Poisson regression. Otherwise, try linear, multiple, or nonlinear regression as needed.
Step 5. Insure your model produces results that are acceptable.

9 Using R

Before we start: setting up the workspace
Before working in *R*, it is necessary to set up the "workspace": the virtual environment in which you can load, manipulate, and analyze data. The code below cleans the workspace, erasing any previous objects or functions; sets the working directory, from which we'll load the data to analyze, and also loads a set of "packages" of useful functions that make data cleaning and analysis easier and faster.

Using the *cor()* command in *R* on the data table:

```
############################
##
## Setting up workspace
##
############################

## Clear previous workspace, if any
rm(list=ls())

## Set working directory
os_detect <- Sys.info()['sysname']
if (os_detect == 'Darwin'){
  setwd('/Users/localadmin/Dropbox/Research/StatsChapter')
}

## Load packages for analysis
pacman::p_load(
  data.table, tidyverse, ggplot2, stargazer, easynls, pscl, pander
)

############################
##
## Reading in the data sets used in this chapter
##
############################

## Read in spring data
spring_data <- read_csv('./Data/01_correlation.csv')

## Parsed with column specification:
## cols(
##   x = col_integer(),
##   y = col_double()
## )

sigacts_data <- read_csv('./Data/06_poisson_sigacts.csv')
```

```
## Parsed with column specification:
## cols(
##   sigacts_2008 = col_integer(),
##   ggi_2008 = col_double(),
##   literacy = col_double(),
##   poverty = col_double()
## )

recovery_data <- read_csv('./Data/02_exponential_decay.csv')

## Parsed with column specification:
## cols(
##   T = col_integer(),
##   Y = col_integer()
## )

shipping_data <- read_csv('./Data/03_sine_regression_shipping.csv')

## Parsed with column specification:
## cols(
##   Month = col_integer(),
##   UsageTons = col_integer()
## )

afghan_data <- read_csv('./Data/04_sine_regression_casualties.csv')

## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Month = col_integer(),
##   Casualties = col_integer()
## )

war_data <- read_csv('./Data/05_bin_logit_conflict.csv')

## Parsed with column specification:
## cols(
##   side_a = col_integer(),
##   cd_pct = col_double()
## )

sigacts_data <- read_csv('./Data/06_poisson_sigacts.csv')

## Parsed with column specification:
## cols(
##   sigacts_2008 = col_integer(),
##   ggi_2008 = col_double(),
##   literacy = col_double(),
```

```
##   poverty = col_double()
## )

alliance_data <- read_csv('./Data/07_bin_logit_alliance.csv')

## Parsed with column specification:
## cols(
##   statea = col_character(),
##   stateb = col_character(),
##   alliance_present = col_integer(),
##   igo_overlap = col_integer()
## )

## Format and subset casualties data
afghan_data <- mutate(
  afghan_data
  , Date = as.Date(paste0(Year, '-', Month, '-', '01'), format = '%Y-%m-%d')
  ) %>% filter(
    Date >= as.Date('2006-01-01')
    , Date <= as.Date('2008-12-01')
  ) %>% mutate(
    DateIndex = 1:36
  )

## Print data as a tibble
print(spring_data)

## # A tibble: 11 × 2
##        x     y
##    <int> <dbl>
## 1     50 0.1000
## 2    100 0.1875
## 3    150 0.2750
## 4    200 0.3250
## 5    250 0.4375
## 6    300 0.4875
## 7    350 0.5675
## 8    400 0.6500
## 9    450 0.7250
## 10   500 0.8000
## 11   550 0.8750
```

9.1 Correlation in R

Using the *cor()* command in *R* on the data table:

```
## Calculate and print correlation matrix
print(cor(spring_data))
```

```
##          x        y
## x 1.0000000 0.9992718
## y 0.9992718 1.0000000
```

9.2 Plotting in R

```
## Generate a plot visualizing the data
spring_cor_plot <- ggplot(
  aes(x = x, y = y)
  , data = spring_data) +
  geom_point() +
  annotate(
    'text'
    , x = 100
    , y = 0.75
    , label = 'Correlation coefficient:\n 0.999272'
    , hjust = 0) +
  ggtitle('Spring data scatterplot') +
  theme_bw()

## Print the plot to console
plot(spring_cor_plot)
```

9.3 Fitting an Ordinary Least-Squares (OLS) Model with Form y y + ϵ to the Spring Data in R

```
## Fit OLS model to the data
spring_model <- lm(
  y ~ x
  , data = spring_data
  )
```

9.4 Correlation Matrix in R

```
## Calculate and print correlation matrix
print(cor(recovery_data))
```

```
##           T         Y
## T  1.0000000 -0.9410528
## Y -0.9410528  1.0000000
```

9.5 Quadratic Regression of Hospital Recovery Data

```
## Generate model
recovery_model2 <- lm(Y ~ T + I(T^2), data = recovery_data)
```

```
## # A tibble: 15 × 6
##        T     Y index predicted  residuals pct_relative_error
##    <int> <int> <int>     <dbl>      <dbl>              <dbl>
## 1      2    54     1 52.460836  1.5391644          2.8503045
## 2      5    50     2 47.640993  2.3590072          4.7180144
## 3      7    45     3 44.575834  0.4241663          0.9425917
## 4     10    37     4 40.200199 -3.2001992         -8.6491871
## 5     14    35     5 34.780614  0.2193857          0.6268164
## 6     19    25     6 28.672445 -3.6724455        -14.6897820
## 7     26    20     7 21.364792 -1.3647924         -6.8239618
## 8     31    16     8 17.033457 -1.0334567         -6.4591042
## 9     34    18     9 14.790022  3.2099781         17.8332119
## 10    38    13    10 12.213370  0.7866302          6.0510012
## 11    45     8    11  8.844363 -0.8443634        -10.5545422
## 12    52    11    12  6.926437  4.0735627         37.0323886
## 13    53     8    13  6.770903  1.2290967         15.3637082
## 14    60     4    14  6.511355 -2.5113548        -62.7838691
## 15    65     6    15  7.214379 -1.2143795        -20.2396576
```

9.6 Prediction

```
## Create a set of hypothetical patient observations with days in the hospital from 1
to 120
patient_days = tibble(T = 1:120)
```

```
## Feed the new data to the model to generate predicted recovery index values
predicted_values = predict(
  recovery_model2
  , newdata = patient_days
  )
```

9.7 Nonlinear Regression

9.7.1 Exponential Decay Modeling of Hospital Recovery Data

```
## Fit NLS model to the data
## Generate model
recovery_model3 <- nls(
  Y ~ a * (exp(b * T))
  , data = recovery_data
  , start = c(
    a=1
    , b=0.05
  )
  , trace = T
)
```

9.7.2 Sinusoidal Regression

The functional form for the sinusoidal model we use here can be written as:

$$Usage = a * sin(b * time + c) + d * time + e$$

This function can be expanded out trigonometrically as:

$$Usage = a * time + b * sin(c * time) + d * cos(c(time)) + e$$

```
## Generate model
shipping_model2 <- nls(
  UsageTons ~ a * Month + b*sin(c*Month) + d*cos(c*Month) + e
  , data = shipping_data
  , start = c(
    a=5
    , b=10
    , c=1
    , d=1
    , e=10
    )
  , trace = T
)
```

9.7.3 Sinusoidal Regression of Afghanistan Casualties

Visualizing data on casualties in Afghanistan between 2006 and 2008 shows an increasing trend overall, and significant seasonal oscillation. Once again, we want to fit a non-linear model that accounts for the oscillation present in the data. We use the same sinusoidal functional form

$$Casualties = a * sin(b * time + c) + d * time + e$$

which as before can be expressed as

$$Casualties = a * time + b * sin(c * time) + d * cos(c * time) + e$$

The logistic model in *R* is treated as one case of a broader range of generalized linear models (GLM), and can be accessed via the conveniently named *glm()* function. Note that because *glm()* implements a wide range of generalized linear models based on the inputs provided, it is necessary for the user to specify both the family of model (binomial) and the link function (logit).

```
## Generate model
war_model <- glm(
  side_a ~ cd_pct
  , data = war_data
  , family = binomial(link = 'logit')
  )
```

9.7.4 Poisson Regression

Visualizing count data in a histogram is a useful way of assessing how the data are distributed.

9.8 Histogram Plot

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

Poisson regression in *R* is also treated as a special case of GLMs, similar to the logistic regression covered in the previous section. As such, it can be implemented using the same *glm()* function, but now specifying the model family as 'Poisson', which tells R to implement a Poisson model. The model we use here can be specified as

$$Y = e^{\beta_0 + \beta_1 GGI + \beta_2 Literacy + \beta_3 Poverty}$$

```
## Generate model
sigacts_model <- glm(
  sigacts_2008 ~ ggi_2008 + literacy + poverty
  , data = sigacts_data
  , family = poisson
)
```

References and Suggested Reading

1. Affi, A., & Azen, S. (1979). *Statistical analysis* (2nd ed., pp. 143–144). London, UK: Academic Press.
2. Devore, J. (2012). *Probability and statistics for engineering and the sciences* (8th ed., pp. 211–217). Belmont, CA: Cengage Publisher.
3. Fox, W. P., & Fowler, C. (1996). Understanding covariance and correlation. *PRIMUS, VI*(3), 235–244.
4. Fox, W. (2012). *Mathematical modeling with maple*. Boston, MA: Cengage Publishers.
5. Fox, W. P. (2011, October–December). Using the EXCEL solver for nonlinear regression. *Computers in Education Journal (COED), 2*(4), 77–86.
6. Fox, W. P. (2012). Issues and importance of "good" starting points for nonlinear regression for mathematical modeling with maple: Basic model fitting to make predictions with oscillating data. *Journal of Computers in Mathematics and Science Teaching., 31*(1), 1–16.
7. Giordano, F., Fox, W., & Horton, S. (2013). *A first course in mathematical modeling* (5th ed.). Boston, MA: Cengage Publishers.
8. Johnson, I. (2012). An introductory handbook on probability, statistics, and excel. Retrieved July 11, 2012, from http://records.viu.ca/~johnstoi/maybe/maybe4.htm.
9. Kreutz, J. (2010). How and When Armed Conflicts End: Introducing the UCDP Conflict Termination Dataset. *Journal of Peace Research 47*(2), 243–250.
10. Melander, E., Oberg, M. & Hall, J. (2006). *The 'New Wars' debate revisited: An empirical evaluation of the atrociousness of 'New Wars'*. Uppsala peace research papers no. 9, department of peace and conflict research. Sweden: Uppsala University. http://www.musik.uu.se/digitalAssets/18/18585_UPRP_No_9.pdf. Accessed 12 Sept 2012.
11. Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed., pp. 531–547). Chicago, Il: Irwin Press.

Student Desertion: What Is and How Can It Be Detected on Time?

Jonathan Vásquez and Jaime Miranda

1 Introduction

Student attrition—or student dropout—is understood as a student's failure in completing an educational program, which could be voluntary or involuntary. The first case means a student decides to stop to participate in the program through a formal quitting process according to the school's procedures, while the involuntary, the second case, refers to an institutional decision of finish the student's participation due to disciplinary or academic reasons [30]. The researchers from different disciplines such as Sociology, Psychology, Economy, History, Economy and recently, Data Mining have manifested interest in investigating student attrition phenomenon. Their motivations are related mainly to the desertion's costs, which can be identified from individual (frustration, financial debts, and future income reduction), institutional (funding reduction, opportunity costs, and low performance in indicators related to accreditation) and national perspectives (qualified worker reduction, benefit losses in the government investment and low performance in human development indicator) [29]. However, high complexity in the student attrition researches keep researchers and educational agents creating new methods and tools in order to reduce this global phenomenon [17], generating research opportunities for disciplines for contributing to the improvement of the management of the desertion.

A National Educational system might be considered as the engine that boosts economic and social growth due to its main objective of providing qualified workers, thus, an effective and efficient system on meeting this objective become important for any nation. Currently, Chile is working in a big change of its educational system.

J. Vásquez · J. Miranda (✉)
Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Av. Diagonal Paraguay 257, 8330015 Santiago, Chile
e-mail: jmiranda@fen.uchile.cl

J. Vásquez
e-mail: jovasque@fen.uchile.cl

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_13

Agents related to education like universities, professors, students, and organizations are participating in this process, due they are part of the system and any change will affect them. In this scene, implementing new mechanisms for reducing student attrition is coming considered as the important part of this reform. In fact, Education Minister started some investigation about desertion, and some of its results showed that student attrition is a non-minor problem for higher-education institutes. In fact, the Research Center of the minister, according to data analysis, estimates that 1 of 2 students leave the higher educational system once getting to enter. Apparently, the efficiency of the Chilean educational system creating qualified worker is similar to getting a face (or seal) when a coin is tossed [6].

In addition to the quantification of the problem of student desertion, it is important to understand why this phenomenon rises up, where, from the social sciences perspective, it means to identify the factors and predictors of desertion. Among the first ones, Pyke and Sheridan [23], from an econometric perspective, used logistical regressions obtaining as results that the financial and permanence time in the program are factors that positively influence in the retention. Ten years later, Sadler [26] applied econometric tools to explain desertion for freshman student of a nursing program. He identified that those students who showed a greater internal relationship with the nurse profession (being nurse) had a higher retention rate than those felt an external relation (doing nurse profession). Lately, thanks to the development of techniques and algorithms of Data Mining (DM) discipline, the application of these techniques in studies related to education started to grow, being the identification of predictors of desertion as one of the most important [22]. For example, Delen [8] identified, by the use of data mining techniques, predictors relate academic success as well as those that indicate if the student received funding such as scholarship or loan. In the Chilean case, research with an implementation of methods for student attrition reduction is focused to identify a different kind of student desertions and their costs [2, 9, 13]. For example, some universities have created academic support programs for the student previously they are admitted and once they are admitted to being part of the university. These methods approach on involuntary desertion—mainly academic reasons—, however, few methods, implemented by Chilean institutions, aim at voluntary dropouts. The challenge of conducting research that helps to manage student attrition becomes more important today, since Chilean higher education is in the process of structural changes, increasing free access to educational institutions, and therefore, more investment of the government funding by taxes of Chilean habitats. Therefore, the generation of new tools that improve the management of resources by reducing desertion will allow the investment to generate benefits for society, which would boost national development. In addition, educational institutions could improve their educational management, be fitting to high-quality educational standards reflected on national and international certifications.

2 Understanding Student Attrition

2.1 Spady and Social Variables of the Individual Insertion

Spady used the suicide principles of Durkheim [10]. These principles establish that the decision of suicide of any individual cannot be explained only by individual factors, but also this is a social phenomenon generated by the breakup of the person with his/her social system explained by the impossibility of integration to the society. Following this logic, Spady established that desertion would be a result of the no integration of the individual with his/her educational environment. In other words, he pointed the environment and family characteristics influence on student's expectations, and therefore, affect his or her social integration with the classmates and the final decision of leaving the program [27].

Figure 1 illustrates that family backgrounds impact directly on the academic potential and the attitude compatibility, interests, and the student's personal disposition to the characteristics of the environment. Both the academic potential and the normative conference affect the academic performance, the intellectual development and the integration of the student with the pairs. Each attribute that defines these factors impact directly to the social integration of the individual, which consequently will define the level of satisfaction hence the commitment with the educational institution. Therefore, every social factor related to family and their pairs will influence the final decision of the student's desertion.

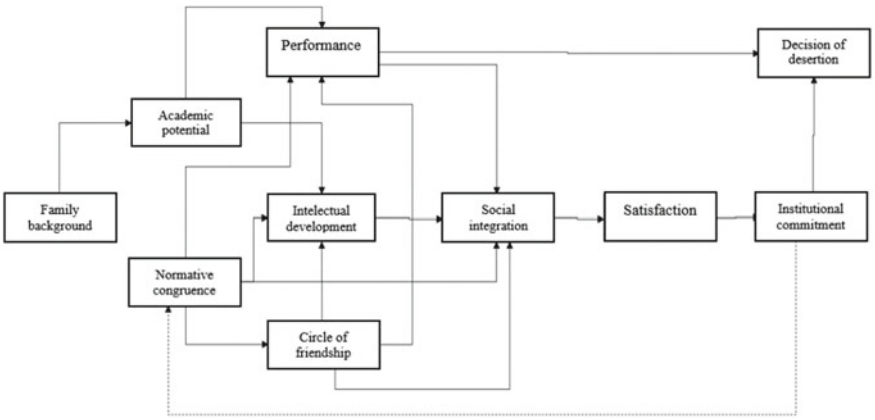


Fig. 1 Factors and their relation according to Spady

2.2 Tinto and the Socio-economic Factors

Tinto and Cullen [30, 31], added the exchange theory to the Spady's model. This theory postulates that human avoids any behavior that involves costs higher than benefits on the relationships, interactions, and emotional states generated by the interaction with their pairs and educational institution. These costs and benefits depend only on socio-economical characteristics of the individual. Under this theory, Tinto suggested that the students would stay in the program as long as the benefits received surpass the effort, dedication, and other personnel costs. Additionally, he established that the commitments of the student with the institution and their personal goals of professional formation are affected by their family background (e.g. socio-cultural level), their personal attributes (e.g. age and gender), and their pre-university academic experience (e.g. university selection exams performance). After a reasonable time enrolled in the program, the student will reevaluate his/her initial commitments according to their social integration and academic performance on the institution, which effects could trigger student desertion if the student notices that the costs are larger than the benefits. In short, Tinto proposed a rational behavior of the student in the continuous-evaluated decision of leaving or stay in the program.

2.3 Bean and the Organizational Factors

Bean [3] explained that the variables related to the student background, such as socio-economical, previous academic performance, and current residency, would influence the determinants of the student's relationship with the educational environment. In short, those students that have academic excellence records at the high school would get better performances at higher school, that it would increase the degree of satisfaction, institutional commitment, and, therefore, higher probabilities of no-desertion as the final decision.

Five years later, Bean joined Metzner [4] and extended the study to non-traditional students. In this new research, they postulated that the socio-demographic variables such as gender, age, and ethnicity are important to the heterogeneity of the body student, and this is important for any research about student desertion.

In summary, the variables identified as important for any study may vary according to institutional context, however, according to theory; we can establish that any research may consider at least three important groups: socio-economic, academic (both pre-university and university) and social (Figs. 2 and 3).

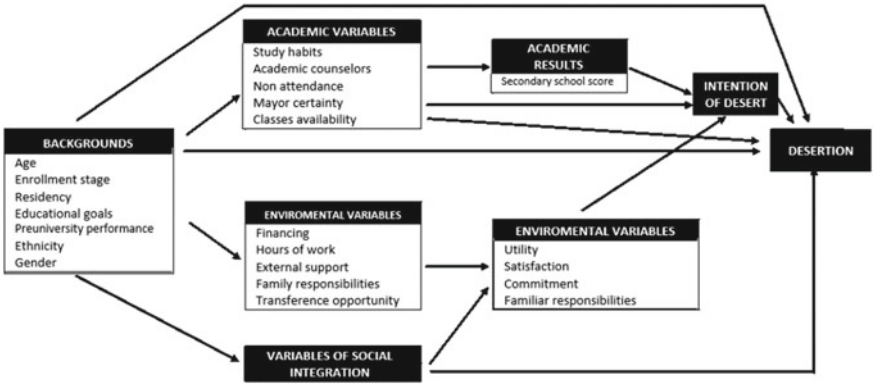


Fig. 2 The relation between the variables suggested by Bean

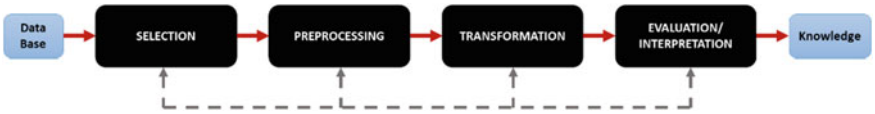


Fig. 3 Knowledge discovery in databases methodology suggested by Fayyad et al.

3 Literature Review

One of the objectives student dropout investigations is, amongst others, the identification of variables that would help to explain desertion, so they could be considered as predictors to detect early the student's decision of leaving the program. Econometric researchers have covered this objective by the identification of statistical significance of variables in the development of regression. For example, Pyke and Sheridan [23] used logistic regressions on a database of 601 postgraduate students, where academic, demographic, and funding were used as potential explanatory variables of desertion. The results showed that the greater the time the student is in the program and the funding awarded, the greater are the chances of graduating. In another research related to undergraduate nursing programs [26], the author compared two evaluation tools used for admission: Grade Point Averages (GPA) and a personal statement essay. According to the statistical analysis, the author found a difference statistically significant in the essay evaluations between deserting and non-deserting students, i.e., those students that completing the program tended to write in the essay their relationship with nursing in a more internal (being nurse) than external way (doing nurse profession). However, he did not find a statistically significant difference in the GPA, unlike other previously published studies did find it [5]. These investigations mostly have helped to identify explanatory variables of the desertion, opening a door for other disciplines, such as data mining, to contribute to the study the generation of predictive models.

The grow technologies' computation capability and reduction-storage costs have facilitated the implementation and development of new Data Mining techniques as solutions in any educational institutes. Indeed, researchers identified the challenge of using DM techniques tools in educational contexts, creating a new discipline named Educational Data Mining (EDM). This, considered as an emerging one, covers the development of techniques, methods, and models for the treatment of unique data for educational management. In the case of student attrition, researchers have faced three common problems [22]. One of them is the generalization of variables since the factors identified as most important by economists are not applicable in any educational context. The second problem is the definition of the temporality of desertion, because dropout reasons may vary in each semester. Finally, a third problem is the imbalance in databases.

3.1 Generalization of Variables

Yu et al. [33] applied three data mining techniques to identify desertion, suggesting that although many researchers had used parametric techniques, such as linear and logistic regression, this new perspective were new and allowed to detect non-linear and non-conventional relations between variables. In their research, features about hours transportation, residence (in or out campus), and ethnicity were considered the most crucial to predicting the early exit of a student. These variables differed from those found in econometric research, which indicated that academic performance at school was the most important predictor of attrition. Additionally, authors concluded variables identified by econometrical investigations were not generalizable for all educational context and called to the implementation of data mining techniques for future investigations. Following this tendency, Delen [8] identified that the variables considered as best predictors for the university context of its investigation were the academic success before and during university as well as the type of funding aid (scholarship or loan). In short, although the econometric studies allow initially to consider an initial set of variables, data mining techniques help to evaluate if these are applicable in different educational contexts, since as we saw previously.

3.2 Temporality and Unbalance

The concentration of desertion's occurrence is not always the same for all educational programs, due to the number of semesters and other specific features of the program (complexity, certification, and etcetera). In fact, it is possible to identify variations for the same kind of programs at amongst institutions from the same city, region, and country. This might imply that occurrence of dropout does not always focus on the same semester for all educational institutions, forcing researchers to identify when students drop out of the study program. For example, Alkhasawneh and Hargraves

[1] formulated a hybrid model with the aim of predicting retention for the first year; while Yu et al. [33] decided to do it for the second and third year. The effect of temporality also generates a difference among nations, since the programs' extension (in semesters) varies, even, amongst countries, being particularly longer in Latin America than in Europe or North America. In the other hand, in relation to the characteristics of the database, researchers have identified an imbalance in the data, being in general much less the number of students who deserted. This generates a problem in the data mining models, causing low precision for the desertion class and high for the retention. In short, it is important to evaluate and apply balancing techniques in order to reduce the problems associated with this kind of databases [8, 28].

3.3 Chilean Context

Díaz [9] did an extensive review of articles and publications related to student desertion. He concluded that there was a low volume of national investigations where data-mining techniques are applied in order to understand the desertion phenomena. In the same way, Himmel [17] claimed for a shortage of national investigations. Since then, study of student desertion has taken much more interest in educational institutions and researchers and they started to run a set investigations in order to understand student desertion and apply different techniques would support retention programs or actions to mitigate the costs of desertion [9, 20, 21].

4 Discovery of the Knowledge on the Database (KDD)

The authors Fayyad, Piatetsky-Shapiro, and Smith proposed a methodology named as Knowledge Discovery in Database (KDD) [11, 12]. They suggested a set of non-trivial activities with the objective of apply techniques correctly with analytic focus and then, increase the probability of projects success. Fayyad and his co-authors indicated that the basic issue approached by the KDD methodology is the data mapping of low level to identify knowledge and patterns; this would allow applications on different areas such as marketing, fraud detection, factoring, telecommunications, medicine, human resources, and education.

4.1 Selection

In this stage, the variables and observations are selected to be used in the data-mining project. The database is explored and analyzed with the goal of identifying variables that fulfill the following requirements: deterministic to the research; none

or almost non-register error, be available in the future if the analysis is made again, and is measurable and collected on time according to the occurrence of the studied phenomena.

The researchers use statistical techniques and technological tools like automatic learning for support the selection of variables. Currently, there is a line of study that looks for generate technological methodologies for supporting variable selection. Reducing dimensions of the final data used in the data-mining project would help to solve the problem of capacity processing, improve the model's performances, and decrease the execution time of the algorithms.

4.2 Pre-processing

Once obtained the database with selected features, then all noise—missing values, outliers, etcetera, and bias must be eliminated. Noise problems can be solved through the replacement and reduction of the observations or variable. Depending on the type of data, the values can be replaced by mode (categorical variables), mean (numerical variables) or by the use of predictive models such as machine learnings. However, it is important to keep in mind that replacement could generate some bias, i.e., the analyst must preprocess carefully the variables.

4.3 Transformation

Some algorithms applied on the data mining stage require that the datasets meet some characteristics, such as only numerical variables. The transformation depends on the kind of variable selected and pre-processed on the previous stages. It is important to identify previously the existence of two kinds of variables (1) Numeric and (2) Categorical. The numeric variables imply that numbers, i.e., age and funding level, represents features and categorical variables capture any information by categories values, i.e., city, and gender. It is important to highlight that categorical variables are generally storage as text but can be eventually a number.

Some machine learning algorithms, as the case of Neural Net, require all variables be numeric and, for a better performance, normalized. The result of this stage generates a set of observations, with variables that are transformed into numbers and normalized according to data mining techniques' requirements.

4.4 Data Mining

The goal of this step is to extract patterns and knowledge previously unknown. The techniques applied can be of clustering, classification, or regression type. The last

two are used mainly to build prediction models: in the case of classification, the tasks refer to the construction of models where observations are categorized on predefined labels; in the other hands, for regression case, activities refer to the use of models to predict variables related to a numerical indicator. The decision of implementing regression or categorization models depends crucially on the objective of data mining project.

4.4.1 Clustering

A clustering algorithm divides a set of observations $\boldsymbol{X}$ at n groups, and each of these groups is featured by a centroid. Currently, there are different clustering algorithms, being the most common k-means [15, 16]. This aims to divide M observations with N dimension into k clusters represented by a centroid. For each partition, the distance amongst observation to the centroid is minimal. Once the sets or partitions are identified, it is possible to create a model for each subset and so improve the predicting performance.

4.4.2 Imbalance Datasets

Imbalance database refers that proportions among different classes are not similar, which creates negative effects on the model's results and performance. The main problem is the model tends to learn more from the majority class than the minority one, or, in other words, the performance is much better in the majority than minority class. In order to solve this problem, it is necessary to evaluate the balancing technique application, i.e., Random Under-Sampling (selecting randomly observations of the majority class for removing from database until getting a balanced database amongst classes) and Random Over-Sampling (selecting randomly observations of minority class to adding the database until getting a balanced database amongst classes) [7].

4.4.3 Machines Learnings

Machines Learnings (ML) concerns to the study of programming computers (machines) in the order they can learn from datasets. ML is considered a branch of artificial intelligence and has been used in a variety of applications, such as text or document classifications, natural language processing, optical character recognition, and lately, educational context [19, 22]. Some examples of these machine learnings are Support Vector Machine, Decisions Tree, Neural Net and Logistic Regressions.

- Support Vector Machine (SVM): Support Vector Machine was introduced by Vapnik and Chervonenkis [32]. Nowadays, from machine learning discipline, SVM is a model of the supervised learning that bases on classification algorithms

and regression analysis. In other words, SVM classifies a set of points in the space by the hyperplane's partition and minimizes the cost of error of classification.

- Decision Trees (DT): Decision Trees is an algorithm considered suggested by Quinlan [24], based on the decision theories to make classification to the databases where algorithms are applied. Quinlan has made big contributions to the algorithms of decision trees, being the most known the C4 and ID3.
- Artificial Neural Net (ANN): ANN was initially introduced as a concept of a neural net by neurologists [18]. Fifteen years later, Rosenblatt [25] introduced the first simple perceptron based on biological neural net concepts, proposing fundaments of an artificial neural net (ANN). Nodes named as neurons compose an ANN, and each node receives a set of entries coming from other nodes and delivers outputs to others.
- Logistic Regression Logistic Regression is a special case of regressions used to predict the result of a categorical dependent variable. As an example, assume that the response variable y takes values 0 or 1, as in the case of desertion (dropout = 1 and not-dropout = 0). According to the postulated by logistic regression, the posterior probability of answer to the conditioned variable of the vector. After, the algorithm identifies the coefficients w of iterative form, usually through the method of maximum likelihood [14].

4.4.4 Classifiers

The learning machines deliver an indicator that shows the probability of a register belonging to a particular class. In some cases, depending on the algorithm, such indicator is reflected on the confidence, calculated by each observation, and can be used to determine classification thresholds, where to all confidences over the b threshold are classified to one class.

As an example, imagine that we have datasets of students and a machine learning is applied, so, after obtained prediction function, we obtain a confidence dropout class for each observation. On a predefined way, once the observations are ordered from the highest [1] to lowest (0) confidence, as a standard, observations with a confidence equal or bigger than 0.5 are classified as a dropout. This threshold could be more or less restrictive, increasing the classification threshold in the case that a bigger restriction is required to catalog a register as dropout and therefore less restriction to the other class no-dropout, or decrease the threshold if it is required to be less restrictive to the dropout class and more restrictive to the other one.

4.5 Interpretation and Evaluation

After Data Mining step, it is necessary to identify whether models are good or bad ones. In this step, evaluation and interpretation, the analyst must dominate the project context; due he/she should evaluate the results, and relate them to studied phenomena.

The performances of implemented models could vary for many factors, such as variables selected, the used algorithms, implementation of an appropriate optimization parameters process, among others. Therefore, it is imperative to use mechanisms to evaluate the performance of each technique in order to identify the best one. In this sense, literature has suggested different metrics to measure the predictive performance of the models. The most commons are the classification error and the accuracy. However, these metrics measure the general performance of models, assuming that classifying improperly any class imply the same errors, this means, have the same cost. Obviously, in some cases, like in the case of student attrition, to classify a student as dropout and finally stay, has not the same cost that classify the same student as not-dropout and finally leave. In short, the performance of the models can be evaluated based on its accuracy and classification cost. To measure both metrics it will be used the matrix of confusion, the tool usually used in classification applications, given the easiness of usage and quality of information delivered.

4.5.1 Confusion Matrix

The confusion matrix is a table that consists mainly of 2 rows and 2 columns—depending on classes number—, with information on the performance of the classification of a classification model. Usually, rows represent the instances predicted by the model, meanwhile, columns do the real instances observed. Additionally, the datasets are separated into two groups: train dataset, which is used for finding the function of the model, and test dataset, which is used for evaluation of the model's performance.

In case classification of two classes, these are named as positive and negative classes. A prediction of the class is obtained for each observation of test dataset by implementing the model generated through a learning process where train dataset was used. Each prediction is compared with the real class, and those observations predicted as positive and effectively were positive are denominated as True Positive (TP), but if they were not, they are evaluated as False Positive (FP). It is a similar case for the negative class, assigned as True Negative (TN) those observations predicted as negative and effectively were negatives, meanwhile, as False Negative (FN) the observations were positive but the model identified them as negative. The next table illustrates a typical confusion matrix (Table 1).

Table 1 Example of the confusion matrix

	True class		
Class prediction		+	−
	+	True positive	False positive
	−	False negative	True negative

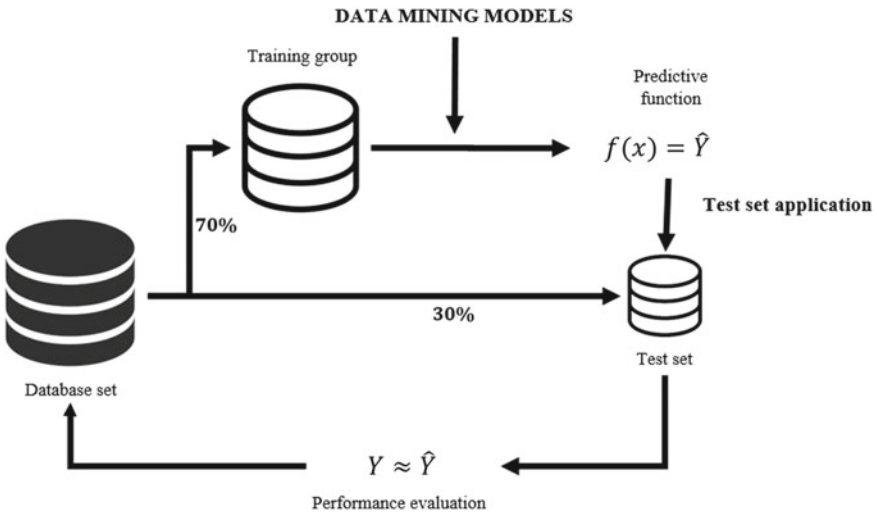


Fig. 4 Representation of cross-validation

4.5.2 Cross-Validation

Cross-Validation is a validation method mainly used to estimate how accurately a model will perform. As it was explained before, this method divided the datasets into two, train and test. This process is applied to k subsets of original datasets, everyone of equal size. One of these subsets is retained as the test data set. The remaining subsets are used as train data set, then the cross-validation method is repeated k times, with each of the k subsets used exactly once as the test data. The final performance evaluation is obtained from the k results of the k iterations. The performance is showed in averaged of these k results, so it is produced just one estimation of performance. The values most accepted for k are 5 or 10. Figure 4 illustrates a cross-validation method.

5 Experiment

5.1 FEN Case

The Facultad de Economía y Negocios (FEN, by its acronym in Spanish), it is the Business School of the University of Chile and it is located at the top ten on Latin-American rankings. The school manages four programs: (1) Commercial Engineering focused on Business (Business and Administration), (2) Commercial Engineering focused on Economics (Economics), (3) Audit, and Information Systems and Management Control Engineering. This study will focus on the last program.

Table 2 Distribution of desertion per entry groups

Year of entry	Non-desert (%)	Voluntary desert (%)
2007	66.0	34.0
2008	74.0	26.0
2009	74.0	26.0
2010	68.9	31.1
2011	59.6	40.4
2012	67.6	32.4
2013	79.8	20.2
2014	77.5	22.5
Total	70.9	29.1

In Chile, there is a national system for the application process to higher-education programs. This consists of a university selection test named University Selection Test (PSU, by its acronym in Spanish), that measures the knowledge of a student in four areas: Verbal, Mathematics, Sciences (Biology, Physics, and Chemistry) and History, Geography, and Social Sciences. Each test evaluates the knowledge of the students on a scale of 150–850 points. The attached universities to this process establish weights for each test, including the grades average obtained in high school, according to the program offered by each institution. The students, weighing the scores, apply by the national system to the programs, ordering these applications by preferences. All vacancies are completed by order of the score according to the weight established by the institutions.

It is important to highlight that each student can apply to up four programs and they are selected only in one. In the case of the Information Systems and Management Control Engineering program, the total of vacancies is at least 100 approx., varying each year according to evaluations and projections made by the school.

Approximately, in the case of Information Systems and Management Control Engineering program, 30% of the total of students that enter in a year, leaves voluntarily its studies, translating into an average of 31 spots lost every year. The variation per year of desertion it is showed in Table 2, where a 20 and 40% of a group of students with the same year of entry, quit the program voluntarily.

The six firsts semesters of the program are crucial to detecting voluntary dropout, because as shown in the Fig. 5, 97% of desertion occurs in the first three years, focusing on the third semester. In short, this study will focus on dropout that occurs during the first six semesters.

5.2 Databases

The dataset is collected from three databases: (1) Educational Administration System (SAD, by the acronym in Spanish), (2) Scholarships and Credits and (3) DEMRE

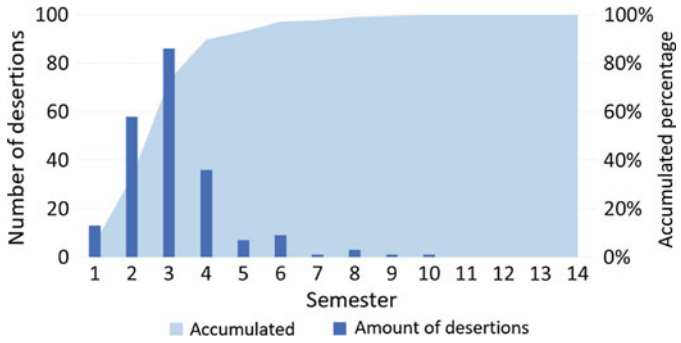


Fig. 5 Desertion behavior per semester. Own making

Table 3 Number of observations by semester

Semester	Desert	Non-desert	Total
Sem 1	13	595	608
Sem 2	58	591	649
Sem 3	86	591	677
Sem 4	36	578	614
Sem 5	7	486	493
Sem 6	9	577	586
General total	209	3,418	3,627

base. The Educational Administration System (SAD) stores and keeps students' information about enrollment, homologation of credits, academic performance, student's requests (such as temporarily stop studying and leave the program), and professor evaluation made by the students. In the other hand, information stored at the Scholarships and Credits is related to funding, as much in amount and funding type (scholarships and credits) for each student that receive each year funding such as Scholarships and Credits from the Ministry of Education and the internally from the University of Chile and FEN. Finally, the database sent annually by the Department of Evaluation, Measurement and Educational Register (DEMRE by the acronym in Spanish), managed by the University of Chile, has socio-demographic information pre-university academic performance, PSU scores and postulation of each student, given at the same moment of inscription for the PSU test.

The total number of observations was 3,627 with a distribution by semester showed in Table 3. The stored information in different databases allows obtaining 44 variables that cover socio-demographic, academic performances, and environmental and funding information of every student:

- Socio-demographic Variables: the number of family members, parent's educational level, number of parents alive, number of member working and studying at the different levels of the Chilean educational system, gender, and others related to work before entering into higher education.

- Performance and Academic Performance Variables: The variables related to academic performance of the student are stored in DEMRE and SAD systems. The data retrieved were academic performance in high school, score on each section of PSU, academic performance each semester (transcriptions and credits).
- Environmental Variables: Data from DEMRE system enable to obtain in-formation high school type (Technical or Scientific-Humanist), educational regime (Male, Female, or Co-educational high school), funding dependency type (Public Funding, the school is financed completely by the government and managed by municipalities; Private-Subsidized, the high school is co-financed by parents and government and managed by privates; Private, the institution does not receive funds from the government and is completely funded by students' parents and managed by privates), and application preferences. Additionally, from SAD systems, we enabled to obtain information related to postponements, voluntary participation on summer semester, and appreciation of the teacher's performance, which is considered as a proxy of satisfaction with the educational institution.
- Funding variables: related to family income, scholarships, and credits that each student receive every year.

5.3 Implementation

From the database, 6 datasets were obtained—each for semesters—, so we looked for six predictive model. We tested blending different techniques so we obtained 48 models for each semester, which generated 288 models in total. The following techniques were combined: (1) Clustering, (2) Unbalance Techniques (ROS, RUS and without unbalancing techniques), (3) Learning Machines (SVM, Neural Net, Decision Trees and Logistic Regression), and (4) Classification Threshold.

The software used was RapidMinner. In order to eliminate any bias and noise (such as missing values or outliers), only full observations were considered for the investigation, due to the low volume of data with these problems. Polynomial variables were transformed into binomial, generating n new columns where n was the number of different unique categories of the variable. From these n new columns, we selected $n - 1$ to avoid multicollinearity. Finally, the numeric attributes were normalized to a scale of 0–1 in order to match the range in every variable.

For clustering, we used the operator called X-means that balances the costs associated with precision and complexity of the model and delivers, as a result, the number of optimal centroids. Then, the operator assigns the observations to one of these centroids. This operator was tested in each semester dataset in order to obtain the best model.

Machine Learnings were applied to each cluster and to the complete dataset as well (without clusters), obtaining the performance of each one and identifying which of all was the best. Additionally, before applying Machine Learning algorithm, we tested balancing techniques (ROS and RUS) and evaluated how much was the improve compare to using imbalance datasets.

Finally, the classification threshold was applied in the testing process of the Machine Learning. Rapid Miner has operators that allow identifying the best threshold given the costs of classification errors to each class. Thus, after obtaining the fitted model in the test process, the confidence delivered by the algorithm is used as an input to the operator that identifies the best threshold, and then, the classifier uses this threshold and classify each observation in order to improve prediction performance.

6 Results

6.1 Best Models

In general, for every 6 semesters, the best models were composed of clustering, no-balancing techniques, SVM machine learning, and a classifier with an optimal threshold. Only in the case of semester 5 and 6, the machine learning in the best model was Logistic Regression.

Comparing each technique, Logistic Regression and SVM machine learnings generated better performances to the models, as well as the classifier with an optimal threshold. In the case of the balance techniques, it was not clear its impact on the performance of the models, because is not always convenient its application.

6.2 Most Important Variables

Analyzing the best models of each semester according to the weights, the most important variables are those related to PSU, academic performance at higher school, professors rating, pre-university academic performance, parents' educational level, number of family members and how many are working, the application preference, participation on summer semesters and funding.

Considering those variables amongst the first quartile with the highest weight given by the models, the most important are PSU, specifically score on the verbal section, followed by the parents' educational level, and professor ratings. It seems that the student's background, mainly academic performance, must be strongly considered by the educational institution's manager of FEN, as well as the satisfaction of the student with the professors. In the other hand, extending the analysis to the second quartile, again, the most important variables were PSU performances, mainly on the tests of language and mathematics. It seems that the national university selection test is a good predictor of desertion (Fig. 6).

Analyzing Fig. 7, we identified that those variables related to family configuration and university performance are the most common amongst the 6 semesters. This is consistent with the models discussed by Spady [3], Tinto and Cullen [27], Bean [30],

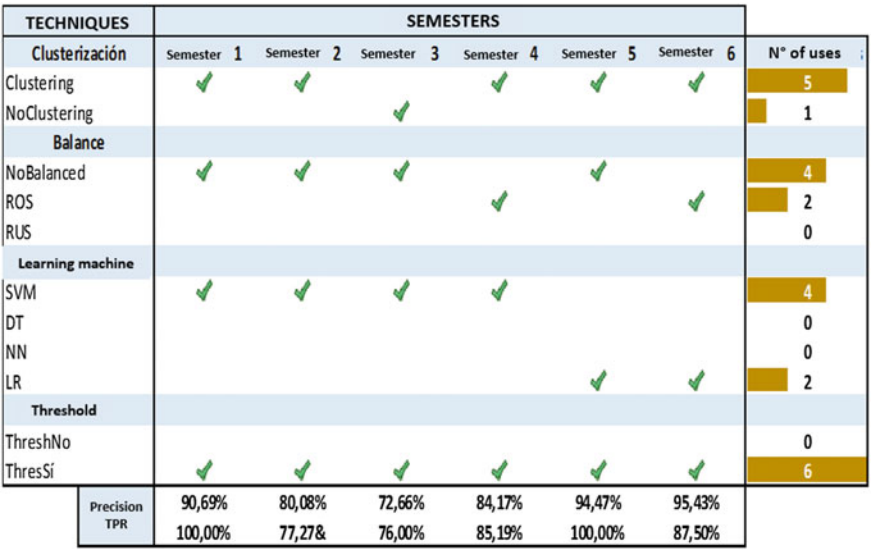


Fig. 6 Performance of the best models by semester

where they proposed that the family backgrounds and the academic performance are primary variables to explain the student desertion.

In summary, for each semester, it is important to evaluate PSU score of the student, followed by the educational level of the parents, university academic performance, funding, and family configuration in order to identify those students wanted to leave the program.

6.3 Deserters and Not-Deserters Profiles

According to the results, it was possible to identify the most important predictors for each semester. The most important for all semesters are the performance in PSU, the number parents alive, teachers rating, and university academic performance. However, the total set of predictors was different for all semesters, i.e., we identified a trend in the set of most important variables while advancing in the semesters. For the first four semesters, predictors related to the student's pre-university features were identified as the most important: specifically, PSU performance in all sections, number of parents alive, teacher rating, and the educational level of parents. In other words, as Spady [27], the variables related to academic potential (pre-university performance), family background, and educational context should impact the student's decision to remain in the program. Managers of educational programs must know these relationships amongst variables, and they could use it as support of selection process, as well as identify which students are potential deserters and implement

Variable	DataBase	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Number of Uses
TeacherRating_Sem	SAD	✓	✓	✓	✓	✓	✓	6
SemGPA	SAD	✓	✓	✓	✓	✓	✓	6
Educational_Level_Father	DEMRE	✓	✓	✓	✓	✓	✓	6
Educational_Level_Mother	DEMRE	✓	✓	✓	✓	✓	✓	6
Number_Members_Family	DEMRE	✓	✓	✓	✓	✓	✓	6
Income_Family	DEMRE	✓	✓	✓	✓	✓	✓	6
Verbal_Score_PSU	DEMRE	✓	✓	✓	✓	✓	✓	6
Math_Score_PSU	DEMRE	✓	✓	✓	✓	✓	✓	6
GPA_HighSchool_Score	DEMRE	✓	✓	✓	✓	✓	✓	6
Funding Level	Scholarships and Credits	✓	✓		✓	✓	✓	5
TeacherRating_Sem_Accum	SAD		✓	✓	✓	✓	✓	5
TeacherRating_Sem_Previous	SAD		✓	✓	✓	✓	✓	5
SemGPA_Accum	SAD		✓	✓	✓	✓	✓	5
SemGPA_Previous	SAD		✓	✓	✓	✓	✓	5
Number_Parents_Alive	DEMRE	✓	✓	✓	✓		✓	5
Verbal_Math_AverageScoring_PSU	DEMRE	✓	✓	✓	✓	✓		5
Number_MembersFam_Studying_HighSchool(4º)Level	DEMRE	✓		✓	✓	✓		4
HighSchool=Private	DEMRE		✓	✓	✓		✓	4
Gender=Male	DEMRE	✓	✓	✓	✓			4
%Failed_Courses_Sem	SAD		✓			✓	✓	3
%Failed_Courses_Sem_Accum	SAD		✓			✓	✓	3
HighSchool=Public	DEMRE		✓			✓	✓	3
%Failed_Courses_Sem_Previous	SAD					✓	✓	2
Postpone_Sem_Accum	SAD					✓	✓	2
Number_MembersFam_Studying_GardenLevel	DEMRE					✓	✓	2
Number_MembersFam_Studying_Others	DEMRE					✓	✓	2
Number_MembersFam_Working	DEMRE					✓	✓	2
Number_Hours_Working(Pre-University)	DEMRE					✓	✓	2
History/Science_Score_PSU	DEMRE					✓	✓	2
School_Type=Technical	DEMRE					✓	✓	2
Educational_Regime=Female	DEMRE		✓			✓		2
Summer_Sem_Accum	SAD						✓	1
Summer_Sem_Previous	SAD				✓			1
Number_MembersFam_Studying_HighSchool(1ºto3º)Level	DEMRE						✓	1
Number_MembersFam_Studying_GardenLevel	DEMRE						✓	1
Number_MembersFam_Studying_HigherLevel	DEMRE						✓	1
Application_Preference	DEMRE					✓		1
Educational_Regime=Male	DEMRE						✓	1
Postpone_Sem	SAD							0
Postpone_Sem_Previous	SAD							0
HighSchool=Private_Subsidized	DEMRE							0
School_Type=Scientific-Humanist	DEMRE							0
Educational_Regime=co-educational	DEMRE							0
Gender=Female	DEMRE							0

Fig. 7 Variables used by best models for each semester

tools to prevent dropouts. Additionally, in the context of the researched program, we advise to extend the Academic Support Program (program that provides tutorials as a reinforcement) to cover more students for the first two years, due, according to results, academic performance of the first two years is considered as an important predictor for the semester five and six.

For the fifth semester, a strong change in the most important variable is identified, i.e., three of the most important variables were only features that are captured once the student entered the university, i.e., the same variables that are important for the first two semesters, are important for the fifth semester.

According to Spady [27], when the student enrolls in an educational program, according to his family and pre-university school background, he/she establishes his/her initial educational objectives and his/her institutional commitment. Later, after a sufficient time in which he/she interacts with its environment, these educa-

tional objectives and institutional commitment are adjusted, so they would trigger a desertion decision. This can be clearly seen in the variables most important by semester, where at the beginning of the program students with female gender, mostly from private-subsidized or municipal schools, with members of the family studying, and with high levels of funding, do not desert. Some of these characteristics, in general, are repeated in the three years, as it is the gender, members studying in another educational institution and the dependency group of the school.

The variables related to the participation of the student with their academic environment begin to take importance, mainly for the second year, where deserters are those students with low accumulated performance and provide a higher evaluation rate to their teachers.

The family configuration usually appears in all semesters, specifically, in desertions profiles. For example, as the higher educational level of at least one of the parents, the greater is the tendency to voluntarily desert by the student. The same happens when the number of members of the family studying in higher school increases. However, this is not the same when the members are studying at another educational level. Eventually, this could show the difficulty of the family in which two members are studying at the university.

7 Discussions and Conclusions

In general, a deserter profile can be described as those students who studied in private high schools, are male, their parents have higher educational levels, high family income, they have family members studying, they provide higher rating evaluation to their teachers until the third semester, and their performance at the PSU is slightly higher. In the case of no-deserters, their profile can be described as those who studied at private-subsidized high schools, are female, their parents have a lower educational level (less than a complete high school), the family incomes are relatively lower, receive higher funding, and PSU scores and high-school GPA are not very high.

The profiles previously mentioned would allow an early identification of students who would decide to stop studying in a specific semester. Additionally, given that academic performance variables are more important in all semesters, educational managers and advisor could pay attention to the academic performance of students, or, develop more supporting academic programs for the students.

As 5 of the 6 best models are composed of clustering techniques, it might be a sign that students tend to group and maybe, the characteristics of his/her group influence in the decision of desert. Indeed, the main variables that establish a difference amongst the groups are those related to family background and academic performance, variables directly related to deserter and non-deserter profiles.

The six models generated would allow the school to predict desertions in the first six semesters. These predictions can be used as an input in order to develop educational policies and reduce desertion rates, such as workshops of professional

contextualization in order to improve institutional commitment and academic or psychological support programs in order to maintain personal objectives.

Acknowledgements Authors would like to appreciate the help provided in the data access for this investigation by Ariel La Paz, Director of Information Systems and Business School, Cesar Ortega, Chair of the Student Records Unit, and Marcia Oyarce, Social Assistant.

References

1. Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in stem disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research, 15,* 35.
2. Barrios, A. (2013). Deserción universitaria en Chile: incidencia del financiamiento y otros factores asociados. Revistacis.
3. Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education, 12,* 155–187.
4. Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55,* 485–540.
5. Byrd, G., Garza, C., & Nieswiadomy, R. (1999). Predictors of successful completion of a baccalaureate nursing program. *Nurse Education, 24,* 33–37.
6. Centros de Estudios MINEDUC. (2012). Serie Evidencias: Deserción en la educación superior en Chile.
7. Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6,* 1–6.
8. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49,* 498–506. https://doi.org/10.1016/j.dss.2010.06.003.
9. Díaz, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estud. Pedagógicos Valdivia, 34,* 65–86.
10. Durkheim, E. (1951). Suicide: A study in sociology (J.A. Spaulding & G. Simpson, Trans.). Glencoe IL Free Press. Work Publ. 1897.
11. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17,* 37.
12. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39,* 27–34.
13. González, L. E., & Uribe, D. (2002). Estimaciones sobre la "repitencia" y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones. Rev. Calid. En Educ. Cons. Super. Educ. Diciembre Del 2002 (Vol. 77).
14. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques. Elsevier.
15. Hartigan, J. A., & Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.
16. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28,* 100–108.
17. Himmel, E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación, 17,* 91–107.
18. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115–133.
19. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.
20. Morales, F., Fuentes, R., Riquielme, S., & Kraemer, H. (2011). Impacto de la intervención del programa de inducción, adaptación y vinculacón a la vida universitaria en la facultad de ciencias empresariales de universidad del Bío Bío. Presented at the ENEFA (pp. 2730–2757).

21. Morales, F., Riquelme, S., Bascuñan, E., & Navarrete, M. (2014). Estudio sobre el éxito académico de estudiantes de ciencias empresariales de la Universidad del Bío-Bío. Presented at the ENEFA.
22. Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41,* 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042.
23. Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education, 23,* 44–64.
24. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1,* 81–106.
25. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386.
26. Sadler, J. (2003). Effectiveness of student admission essays in identifying attrition. *Nurse Education Today, 23,* 620–627. https://doi.org/10.1016/S0260-6917(03)00112-6.
27. Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1,* 64–85.
28. Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41,* 321–330. https://doi.org/10.1016/j.eswa.2013.07.046.
29. Tinto, V. (2007). Taking student retention seriously. Syracuse University.
30. Tinto, V., & Cullen, J. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89–125. https://doi.org/10.3102/00346543045001089.
31. Tinto, V., & Cullen, J. (1973). Dropout in higher education: A review and theoretical synthesis of recent research.
32. Vapnik, V., & Chervonenkis, A. (1964). *A note on one class of perceptrons* (p. 25). Remote Control: Autom.
33. Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8,* 307–325.

Bioelectric Potential Plant

Imam Tahyudin and Hidetaka Nambo

1 Introduction

In Japan, the aging society is the very big problem. In 2014, a publication of the aging society published by the Japanese cabinet office, announced in October 2010 and October 2013 which there are 23% and 25.1% of the elderly population respectively [1]. Their average age is more than 65 years. This condition is the highest proportion in the world [1, 2].

Based on the research of Nomura et al. [1] the condition of the elderly is mapped into two groups: the elderly who live with their families and who live alone. Based on data from samples taken in one of the major provinces in Japan, Kyoto, mention that the number of the second group in 1990 are 43.416 (13.3%) then in 2010 has increased by 110.366 (18.2%).

These conditions lead to various problems one of which is the death that do not known by others, whether caused by accidents in their home or other factors such as murder. Based on the same research, the deaths caused by accidents in the home because it was not helped as much as 12.5% [1].

This reality makes the increasing demand of indoor monitoring. One of the measures being initiated is to examine the installation of CCTV cameras. This camera can monitor that accidents and immediately helped by a neighbor or an authorized officer. However, this solution does not accept because of privacy concerns. Moreover, the use of infrared sensors tested to solve this problem. Despite of the results

I. Tahyudin (✉) · H. Nambo
Artificial Intelligence Laboratory, Division of Electrical Engineering and Computer Science, Graduate School of Natural Science and Technology, Kanazawa University, Kakuma, Kanazawa, Ishikawa 920-1192, Japan
e-mail: imam@blitz.ec.t.kanazawa-u.ac.jp

I. Tahyudin
Department of Information System, STMIK AMIKOM Purwokerto, Jl. Letjendpol Soemarto, Purwokerto, Central Java 35127, Indonesia

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_14

were pretty good but it is high cost because it requires many sensor cells [3]. Then, the other solutions have been tried by using the sense of odor but the results are not so good because there are much noisy when the data records [3]. Regarding this problem, we proposed a solution which by using bioelectric potential sensor. This is able to be used for detecting human behavior and friendly to use in private area. In addition, the cost is achievable.

Therefore, this study outlines are explained about the living plant monitoring by various methods such as machine learning and deep learning. Finally, the last session is conclusion of discussion.

2 Monitoring by Bioelectric Potential Plant

Plant of bioelectric potential generates a low electrical signal because of the plant activities such as photosynthesis and transpiration. Furthermore, the electrical signal will change because of environmental factors such as temperature, humidity and human behavior. The use of bioelectric potential plant could be the solution for monitoring human activities in private area like bath room or bed room. Moreover, it is low cost and it could be a healing media because it produces an oxygen to reduce stress and gives feeling fresh [4–8].

Based on research Hirobayashi et al. [9], states that human activities like stepping around the plants produce a strong correlation with changes the signal by using plant bio-electrical potential. Another study conducted by Nomura et al. [4], Shimbo and Oyabu [10], explaining that human behavior such as moving, walking, communicating and opening the door can be distinguished using bio electric potential. Furthermore, research conducted by Jin [3], utilizing bio electric potential plant to determine the distance of human to the plant by using Artificial Neural Network (ANN). Furthermore, Nambo [5], Nambo and Kimura [6], Nambo [7], Nambo and Kimura [8], conducted research to determine the location of one's position around plant of bioelectric potential. They were using method of classification, J48 algorithm, multi-layer perceptron (MLP) and deep learning method such as CNN (Convolution Neural Network).

2.1 Measurement of Potential

Plant type of this experiment is photos (Fig. 1a) which its leaves are put on two electrodes (Fig. 1b). To perform measurements is using a data logger (Fig. 2). Specification of data logger used is GRAPHTEC GL400-4. It measures the low voltage at an average altitude of sampling, approximately 512 Hz. When there is a human activity like walking, the signal is responding on data logger and then the signal results are stored on a PC in real time via the local network (Fig. 3).

(a) Photos (b) electrodes

Fig. 1 Bioelectric potential planta

Fig. 2 Data leger

2.2 The Example of Signal from Bioelectric Potential Plant

The respond of bioelectric potential plant is different depend on the distance. Response is stronger when a person is more near the plant. We can see the differentiation of response signal like the Figs below [3]:

a. When the distance is 0.5 m
b. When the distance is 1.0 m
c. When the distance is 1.5 m

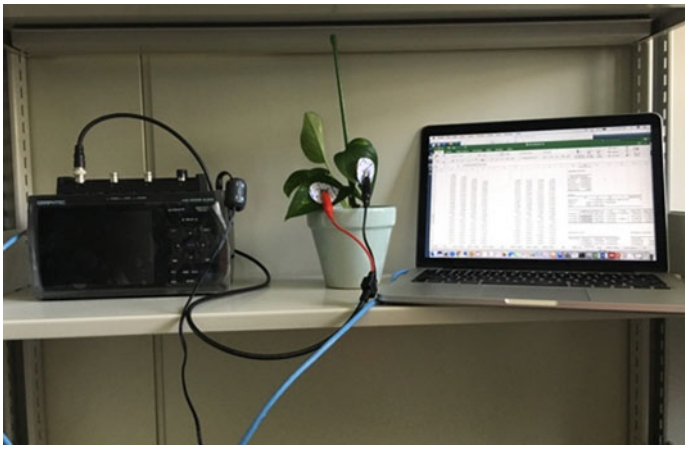


Fig. 3 The measurement processes

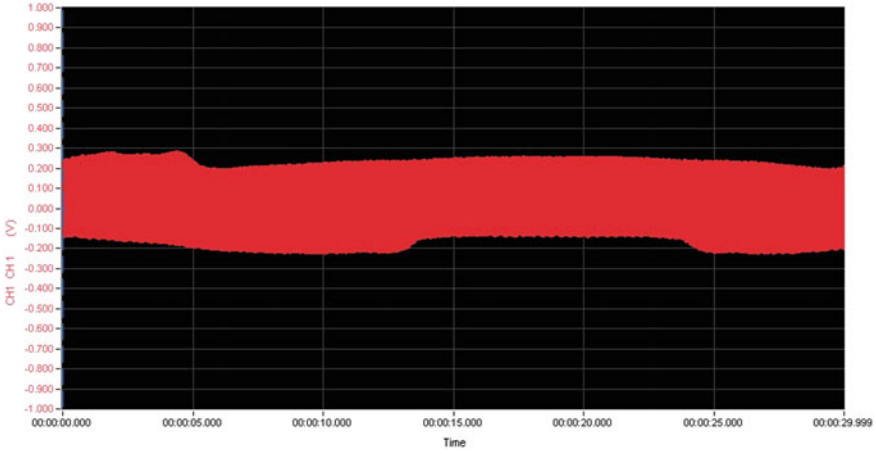


Fig. 4 A plant bioelectric potential when no one is near the plant

Therefore, the response of bioelectric potential is proportional to distance. Base on the experiment if the distance is near so the response is stronger and conversely, if the distance is far so the response is weaker. In addition, this property can be used for human sensor based on distance [3] (Figs. 4, 5 and 6).

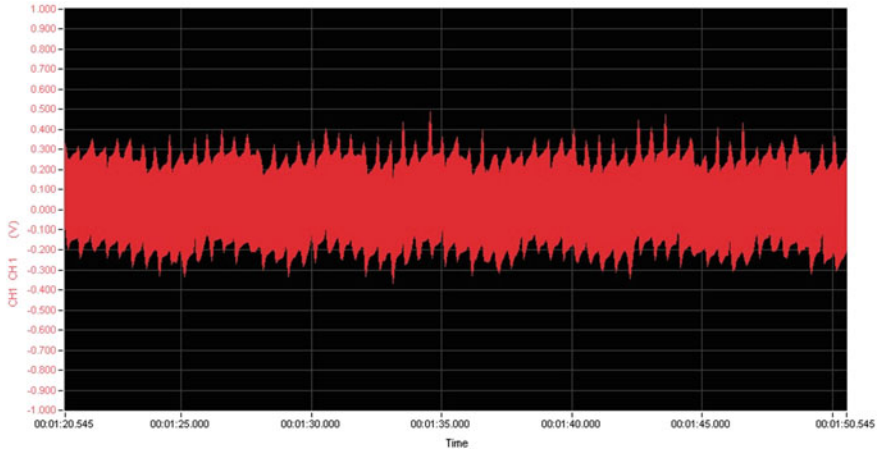


Fig. 5 A plant bioelectric potential when a person is stepping near the plant

3 Monitoring of Bioelectric Potential Plant with Machine Learning

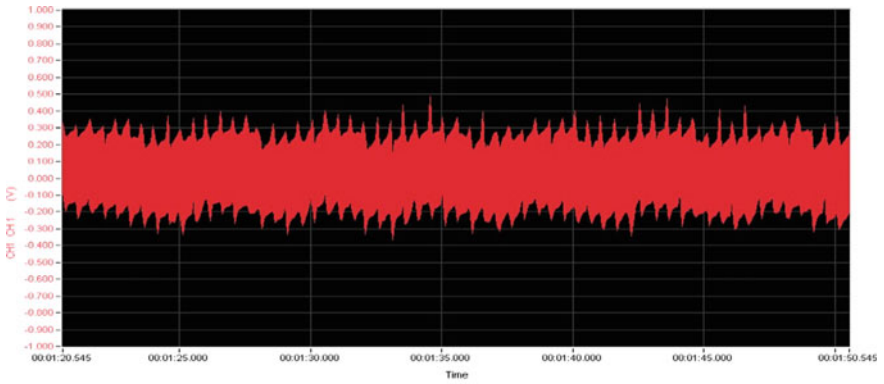
3.1 Proposed System

We would like to propose a method for analyze bioelectric potential plant by using machine learning method (Decision Tree, J48) [5]. This method has aim to estimate the position of person. We make a rule model which the target variables are the existence of person around the plant and the condition when no person in a room. After that, the model is used for examining the testing dataset (Fig. 7).

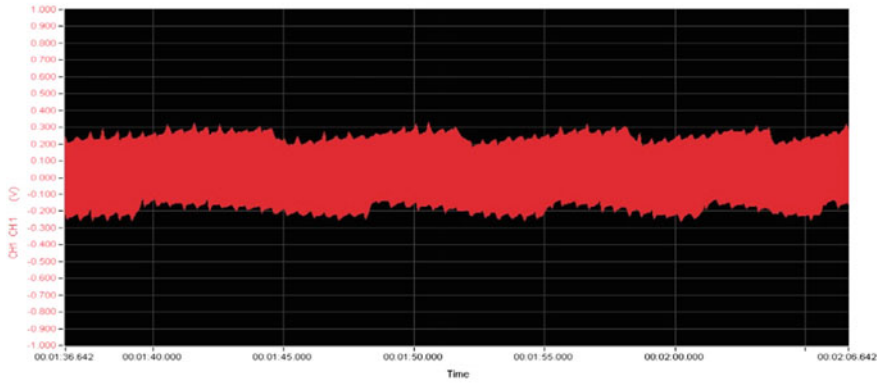
In first, we will explain about data preprocessing method. Bioelectric Potential is continuous data. In our proposed method, we use 512 sample points in every 32 points. Later, this 512 points-data is considered as 1 unit of the process. The description is seen in Fig. 8.

Next, the raw dataset is calculated for obtaining parameters; cepstrum coefficients, average, minimum and maximum value. Cepstrum is the result of the inverse FFT (Fast Fourier Transform) which is the power spectrum of the target signal. This has purpose to get the detail target signal for excellent analysis. The procedure to obtain cepstrum coefficient is below:

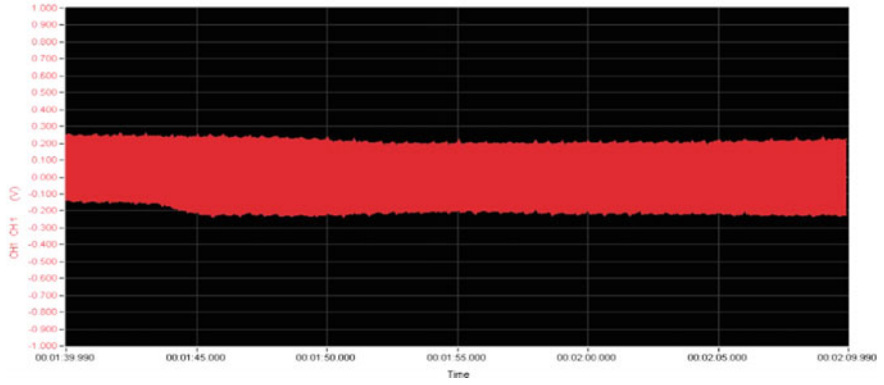
1. Determination of S(t) as a potential signal
2. Determination of power spectrum $S(\omega)$ using FFT of S(t) and square calculation
3. The calculation of the value of log $|S(\omega)|$
4. Determination of Cepstrum coefficients value by calculating the inverse FFT of log $|S(\omega)|$



(a) When the distance is 0.5 m



(b) When the distance is 1.0 m



(c) When the distance is 1.5 m

Fig. 6 The signal responses base on distances

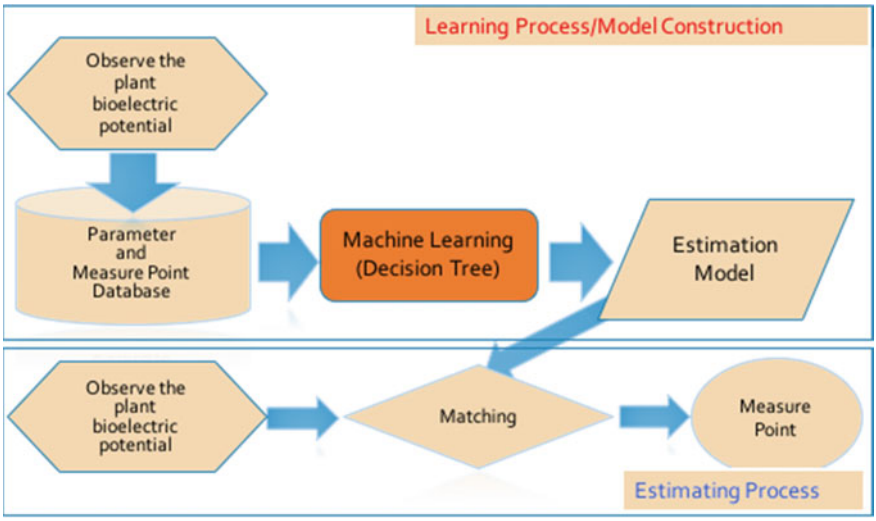


Fig. 7 Bioelectric potential analysis map

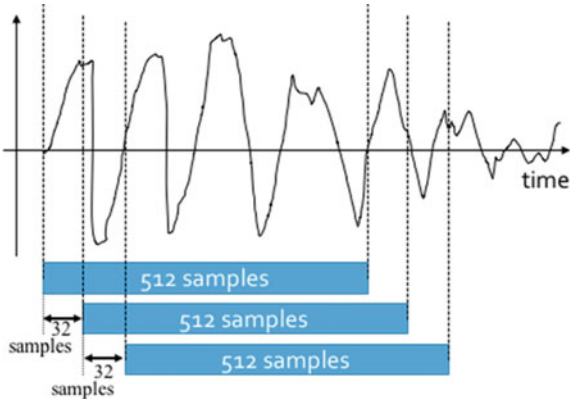


Fig. 8 Determining sample of bioelectric potential data

Based on the calculation, we got 29 parameters which are 26 cepstrum coefficients, maximum value, minimum value, and average from each signal. Totally, we obtain 319 parameters for 11 blocks (Fig. 9).

3.2 Machine Learning Process

After obtaining many instances from pre-data processing, we make estimation model usik decision tree method (J48 algorithm). The number of variables are 319 variables. Observed location is the target variable. We use weka to make the model. See Fig. 10.

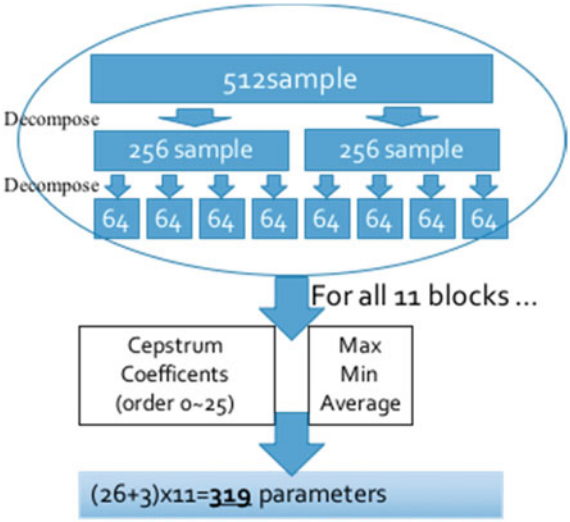


Fig. 9 The parameter number of bioelectric potential plant

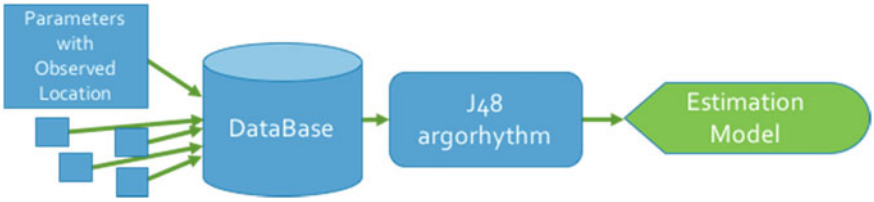


Fig. 10 Machine learning process

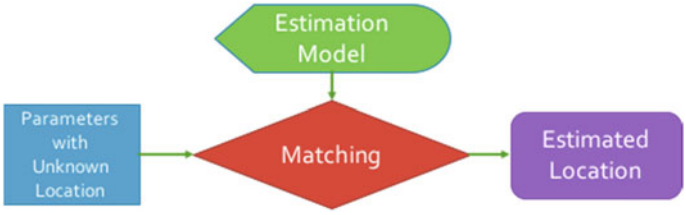


Fig. 11 Estimating process

3.3 Estimating Process

We provide the data testing the same as previous process. By using the estimation model, matching the pre-processed observed dataset (testing dataset). As a result, we obtain estimated locations for each instance (Fig. 11).

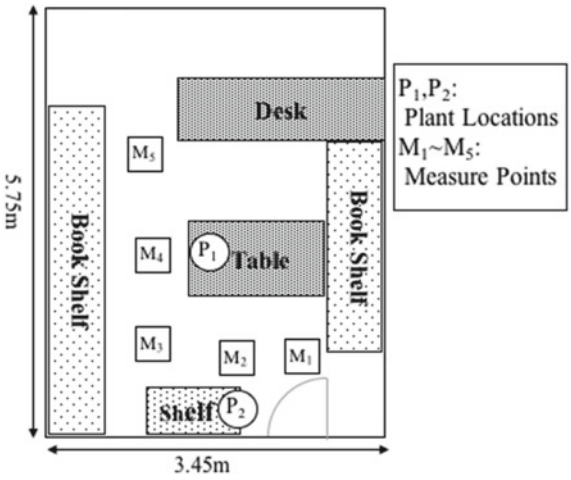


Fig. 12 Experiment environment

3.4 Experiments

The experiment is conducted in a room 5.75 m × 3.45 m in five object position M1–M5 and using two plants, P1 and P2. A person is walking seven times around each object for 30 s. Totally, 210 s observations for each point are conducted. Two plants will respond that human action and the graph changes were seen on the monitor of a data logger. The result of spectrum signal is recorded and is saved in PC (Fig. 2). In addition, when no one in the room, potential is observed for 30 s (point none) (Fig. 12).

3.5 Validation of the Model

Confirm our method and models, we check our model by 10 folding cross validations. Nine of ten instances are used for training dataset and the remain one of ten is used for testing dataset. After we obtain the estimation model from training dataset and we already conducted matching process, see the model validation. The validation process is by changing the test data for each block, test is repeated. Finally, average of 10 tests is evaluation of the model (Fig. 13).

3.6 Result

This experiment uses 10 folds cross validation. Tables 1 and 2 show the model from plant 1 and plant 2. These tables represent input instances measured and the output an

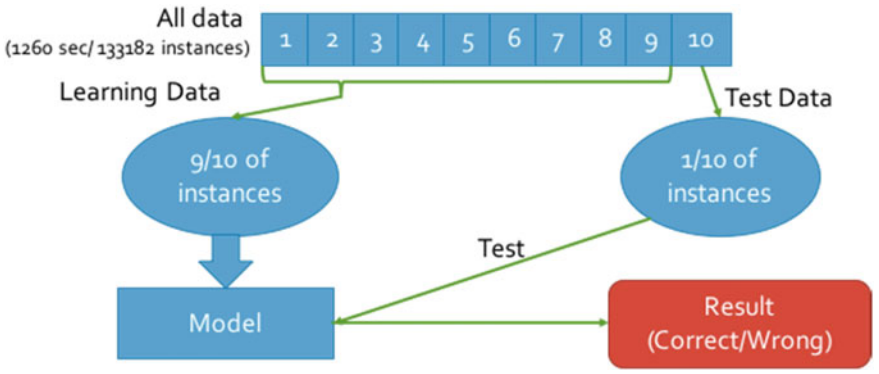


Fig. 13 Validation process

Table 1 Output of matching process from plant 1

Input	Output					
	Ml	M2	M3	M4	M5	None
Ml	2608	507	2	0	14	40
M2	505	2560	21	0	76	9
M3	1	12	2685	6	467	0
M4	0	0	5	3161	5	0
M5	32	73	494	4	2561	7
None	38	14	0	0	3	3116

Table 2 Output of matching process from plant 2

Input	Output					
	Ml	M2	M3	M4	M5	None
Ml	2854	49	265	2	0	1
M2	90	3078	3	0	0	0
M3	232	1	2897	14	10	17
M4	1	0	16	2654	407	93
M5	3	0	8	430	2408	322
None	3	0	11	88	330	2739

estimated point from some object observations, M1–M5 and condition when no one in a room. Both of tables show almost the same as between input and output. Oven thought, there is some misclassification such as M1 and M2 in Table 2. However, the accuracy is fairly high, their F-measure are 0.877 and 0.874 for plant and plant 2 respectively (Figs. 14, 15, 16 and 17).

The property of bioelectric potential plant can be used for estimation of location in the room. Estimation accuracy is high enough.

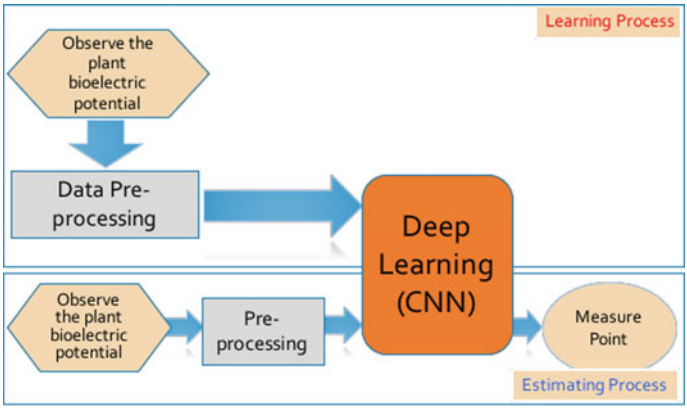


Fig. 14 The process flow of the estimation method

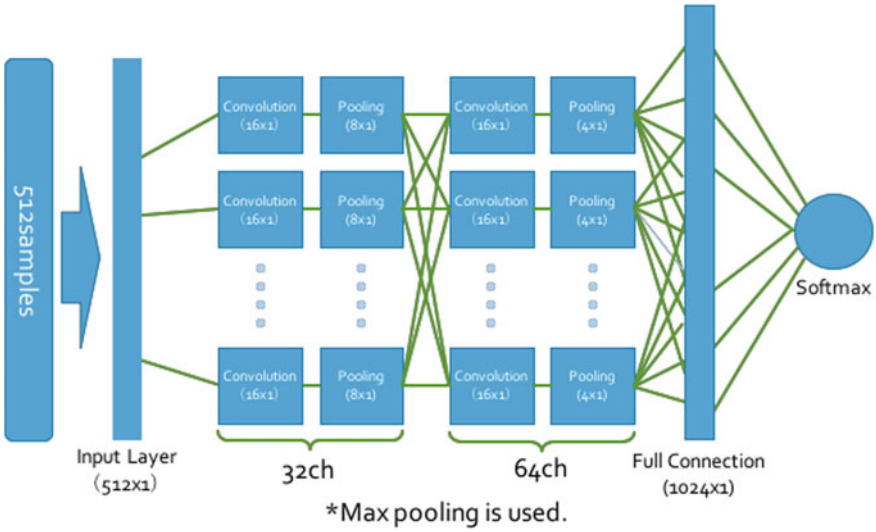


Fig. 15 CNN method with 2 times convolution

4 Monitoring by Bioelectric Potential Plant with Deep Learning

In this study, we propose a method to estimate the location using Deep Learning techniques. In concrete, many feature values are extracted from an observed potential in previous method. However, in this study, observed potential are directly used for estimation [7].

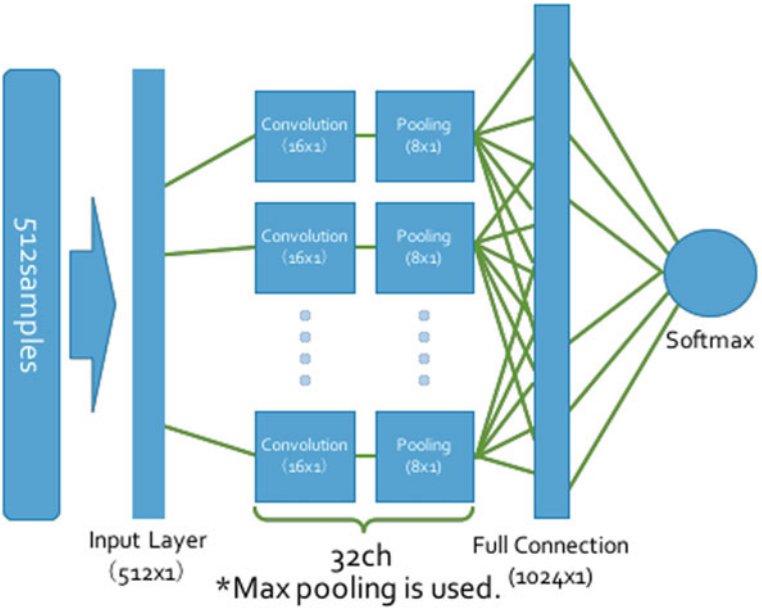
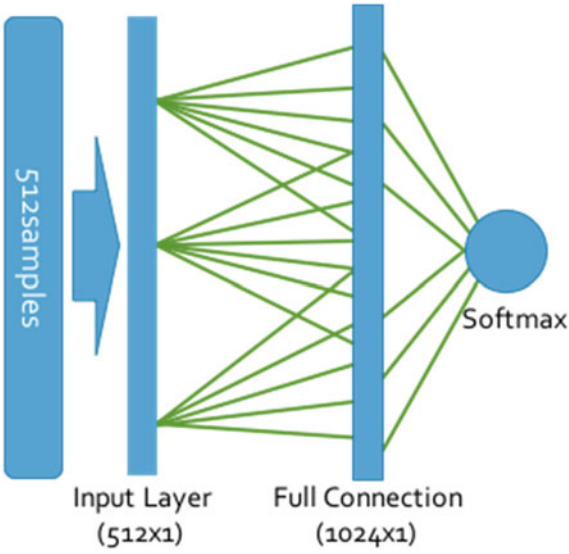


Fig. 16 CNN method with 1 times convolution

Fig. 17 MLNN method



For estimation, we used convolutional neural network. The network accepts 512 points of potential and classifies 6 classes that represents which measure point the person stays or out of room.

4.1 Proposed Method

For data pre-processing is the same as previous method. This proposes method CNN using 2 times convolution/pooling. The result is compared with CNN with 1 time convolution, multilayer neural network (MLNN), and decision tree J48.

Deep Learning Methods

1. CNN using 2 times convolution/pooling
2. CNN with 1 time convolution
3. MLNN.

The comparison of Deep learning with the previous method

Previous Method

- Features are selected by try and error
- It takes a lot of cost and time.

Deep Learning

- We don't care about the feature value
- There are many parameters
- Time consuming.

4.2 Experiment

The experiment environment is the same as previous method (Decision tree J48). There are two plants (P1 and P2) and five point observations (M1–M5). The process is by observing the object for 30 s while person steps around the measurements point Mi, eight times and additionally when no one in the room. Therefore, totally 240 s observations for each point are conducted.

4.3 Training and Evaluation

Training data is extracted from observed potential. There are 453 instances which are extracted from 1 observation (30 s). Observed data of last 1 time observation from 8 observations is used as test data. Totally there are 2718 instances are obtained. And the rest of them, 7 of 8 observations, are used as training data. Totally, there are 19,027 instances are obtained from 5 measure points and 1 absence data.

Training is repeated 20,000 times. 512 instances are randomly selected from training dataset, then network is trained. After that, we evaluated distinction accuracy for test data after 20,000 times training.

Table 3 The comparison of the accuracy from some methods

Method	Plant 1 (%)	Plant 2 (%)
J48 (decision tree)	84.3	71.0
MLNN	58.9	61.4
1 time conv./pool.	31.4	30.2
2 time conv./pool. (proposed method)	86.4	85.4

4.4 Result

Based on the Table 3, we see that accuracy value of CNN with 2 convolution/pooling is higher than others. Its accuracy from both plant 1 and plant 2 are greater than 85%. The second position is by using decision tree method (J48) with the average accuracy is greater than 75.5%. The next position is by using MLNN (the average accuracy is around 60%) and CNN with 1 time convolution (the average accuracy is approximately 30%) respectively.

5 Conclusion

The indoor monitoring is a promise research. The bioelectric potential plant is the recent solution as biological monitoring. There are many research have performed using this topic. They presented good result for various aims. For instances, for detecting human walking, stepping and etc., for determining the human distance with plant, to detect the person position, and many more. On the other hand, this topic is still having problems like accuracy, distance, reproducibility, and so on. Hence, this topic is still developed by trying many methods such as classification method (decision tree J48), machine learning (ANN, MLP), and deep learning (CNN). For the future will try by association analysis and time series approaches.

References

1. Nomura, M., Mclean, S., Miyamori, D., et al. (2016). Science and justice isolation and unnatural death of elderly people in the aging Japanese society. *Science & Justice, 56,* 80–83. https://doi.org/10.1016/j.scijus.2015.12.003.
2. Chen, B.K., Jalal, H., Hashimoto, H. et al. (2016). Forecasting trends in disability in a super-aging society: Adapting the future elderly model to Japan. *Journal of Economy Ageing* 1–10. https://doi.org/10.1016/j.jeoa.2016.06.001.
3. Jin, X. (2014). Recognition of the distance between plant and human by plant bioelectric potential. *APIEMS* 602–606.

4. Nomura, K., Nambo, H., & Kimura, H. (2014). Development of basic human behaviors cognitive system using plant bioelectric potential. *EEJ Transactions on Sensors and Micromachines, 134,* 206–211. https://doi.org/10.1541/ieejsmas.134.206.
5. Nambo, H. (2015). A study on the estimation method of the resident's location using the plant bioelectric potential. *APIEMS* 1896–1900.
6. Nambo, H., & Kimura, H. (2016). Estimation of resident's location in indoor environment using bioelectric potential of living plants. *Sensors and Materials, 28,* 369–378.
7. Nambo, H. (2017). Development of a human sensor using living plant and bioelectric potential. *APIEMS* 19–22.
8. Nambo, H., & Kimura, H. (2017). Development of the estimation method of resident's location using bioelectric potential of living plants and knowledge of indoor bookshelf. In *Proceedings of the Tenth International Conference on Management Science and Engineering Management*.
9. Hirobayashi, S., Tamura, Y., Member, S., & Yamabuchi, T. (2007). Monitoring of human activity using plant bioelectric potential. *IEEJ Transactions on Sensors and Micromachines, 127,* 258–259.
10. Shimbo, T., & Oyabu, T. (2004). Statistical analysis of plant bioelectric potential for communication with humans. *IEEJ Transactions on Sensors and Micromachines, 124,* 470–475.

False Alarms Management by Data Science

Ana María Peco Chacón and Fausto Pedro García Márquez

1 Introduction

The industrial standard ISA-18.2 (2009) [1] determines that "an alarm system is the collection of hardware and software that detects an alarm state, this communicates the indication of that state to operators, and it records changes in the alarm state". Most of business have modern computerized monitoring system to control safety and efficiency the alarms.

Alarms can be also defined as:

- A false alarm is an alarm that is reported when there is no fault.
- A nuisance alarm occurs when it is true but redundant, i.e. the operator receives more than one alert about that alarm.
- A missed alarm is the opposite of a false alarm, it occurs when there is a fault in the system and no alarm has been activated.
- A chattering alarm performs many transitions between normal and abnormal state, it continuously crosses the alarm limit thresholds.

Chattering alarms and alarm flood could be confused, however they are not the same. Alarm flood is when the operator receives many alarms in a short period of time. The alarm flood can be caused by the correlation between variables and, therefore, several alerts are triggered at the same time. The chattering alarm is a single alarm that continuously crosses the alarm threshold, and it generates many alerts. The chattering alarms are defined as those alarms that repeat more than 3 times in a minute according ISA-18.2 (2009) standard [1].

A. M. Peco Chacón (✉) · F. P. García Márquez
Ingenium Reseach Group, University of Castilla-La Mancha, Ciudad Real, Spain
e-mail: AnaMaria.Peco@uclm.es

F. P. García Márquez
e-mail: FaustoPedro.Garcia@uclm.es

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_15

Detection delay is also a key concept. It occurs when the alarms are not activated instantly when the failure occurs. It can occur by the same delay caused by the system itself (deadband, delay-timer, etc.).

In all industrial systems, there are several sensors and actuators for detecting and controlling possible faults. These components can create false alarms, therefore, the control system will be inefficient and the performance will be reduced [2, 3].

Fault detection is an important research area, from the academic and industrial point of view. Numerous methods of detection and control have been designed and developed for fault detection. Some systems can prioritise depending on the gravity of the alarm. When the alarm is triggered, the operator must acknowledge it, to understand it and to know the cause of the alarm, in order to assess its significance and to act to return the operation to its normal state. According to Engineering Equipment and Materials Users Association (EEMUA, 2007) [4], for an operator to respond adequately to an alarm, he must dedicate 10 min to it, i.e. he should not receive more than 6 alarms per hour for a correct operation of the system.

Currently operators receive a large number of alarms, sometimes more than they need or can handle (alarm flooding). It can distract the operator until critical alarms are ignored. Therefore, some operators are reluctant to the monitoring control system. If the system causes many false positives alarms, then it must be redesigned to reduce the number of false and annoying alarms. According to the references [2, 3], most of the alarms received by industrial plant operators are false. There are several methods to improve the alarm system, such as multivariable data analysis or the use of filters, the most important of which are discussed in this chapter.

A few decades ago, only a few selected variables could be controlled. They had to be important for the proper performance of the system and control the quality of the process, because of the alarms were difficult to implement. Each alarm had to be connected by a wire from the sensor to the control room, and it had a high cost. In addition, the control room had limited space and it had to contain numerous control devices. For these reasons, the alarms had to be well designed, to be considered reliable and to guide the operators as it was a good indicator of the correct functioning of the production system.

Nowadays, due to the development of hardware and software, a large number of alarms can be implemented at low cost. Many process variables can be measured and stored in databases. The alarm system communicates with the operator by the Human Machine Interface (HMI) or an Annunciator Panel. Many variables are continuously available in the operator panel for monitoring control. This leads to many alarms, some of them false, chattering or nuisance alarms. Hollifield et al. claim that chattering alarms are the most common type of alarm, where they found 70% of all alarms [5].

Table 1 summarized the number of alarms produced in 39 industrial plants. The alarms have been classified according to their industrial sector and time of occurrence.

According to Walker et al. [7], the U.S. business loses $13 billion a year due to improper use of alarms. Nevertheless, the costs generated by false alarms are difficult to quantify in the world, but this are estimated to be billions of dollars every year. The unnecessary stoppages cause a significant loss of production. For this reason,

Table 1 Performance metrics of industrial alarm system, study carried out in 39 industrial plants [6]

	EEMUA	Oil and Gas	Petrochemical	Power	Other
Average alarms per day	144	1200	1500	2000	900
Peak alarms per 10 min	10	50	100	65	35
Average alarms per 10 min	1	6	9	8	5

the alarm systems should prevent the damage to equipment, downtime and reduced production. Furthermore, the process systems should be controlled to improve the efficiency, availability quality and reliability of the production process [8–10].

There is a great deal of research on alarm management, but the behaviour of the operator in the event of an anomaly is rarely studied. Hu et al. analysed the actions of the operators in response to univariate alarms [11]. The alarm system must clearly and precisely indicate to operators which processes require further attention. They conclude that the operators must regulate the control devices to solve the anomalies of the process, therefore, they can suppress an alarm to temporarily ignore it, in the case of annoying alarms, or they can change the state of the device if an alarm has occurred.

The alarm systems should be designed to help operators to regulate processes and to manage anomalies. Several guides have been written such as ISA and EEMUA for the design, implementation and maintenance of alarm systems [1, 4]. Alarms must be used to ensure the safety of alarm systems and processes. The actions of the operators will depend on the severity of the faults or anomalies that must be announced by alarms. The alerts can be visual or audible.

Monitoring systems are essential to ensure the reliability of the operation of industrial systems [12, 13]. Future hybrid methods provide more robust models in modern and complex installations. Research aims to reduce large production losses and high repair costs due to an inadequate alarm system [14–16].

Many problems are involved in alarm systems, Izadi et al. shown the most common causes [17]: improperly designed alarms, mis calibrated equipment, oscillations in general, changes in status during switching off or on are not taken into account, noise and/or outliers are not considered.

The objective of this chapter is to illustrate the methods and techniques used in several sectors to implement an optimal alarm system. The aim is to obtain: higher quality, higher performance, lower production costs, reduce breakdowns and make the processes safer.

2 Confusion Matrix

An optimal alarm system provides the necessary tools for operators to detect faults and take corrective action to return the process to normal condition. In practice, the alarm control system may be faulty or poorly calibrated, therefore, it will not give correct results. A missed alarm is set when the value of the variable suffers a deviation, e.g. surpass threshold, but the system does not detect it. The opposite case is a false alarm, when the system generates an alarm, although it has not actually occurred, also called as false positive.

Signals can give these two types of errors due to threshold selection. If the threshold setting is very strict to avoid the probability of a missed alarm, this will make the system more sensitive to random noise and the transient deviations and it will lead to more false alarms. On the other hand, if we increase the threshold, the number of false alarms will decrease at the cost of producing more missed alarms. The selection of the threshold is therefore decisive for system reliability. In the majority of cases, missed alarms are considered more important than false alarms, because its consequences may be greater. A basic tool to visualize false alarms in contrast to missed alarms is the confusion matrix.

A confusion matrix, or also called a contingency table, is an evaluation tool for categorical statistical data [18]. The table determines whether the value supplied by the alarm system matches the actual value. The rows of the matrix are the response alarm system and the columns are the actual values. There are 4 possible cases: if the classifier is positive and the system indicates alarm is true positive (TP), if an alarm has not occurred but the system classifies it as such, it is false positive (FP), otherwise an alarm has occurred, and the system does not identify it therefore will be false negative (FN), or missed alarm. Finally, it can happen that no alarm occurs and the negative system, therefore it is true negative (TN). The main diagonal values show when the system has acted correctly. However, the values of the other diagonal show when an error has occurred (Table 2).

The following rates are obtained from the confusion matrix.

$$FP\ rate = \frac{FP}{N} = \frac{negatives\ incorrectly\ classified}{Total\ negatives}$$

$$Precision = \frac{TP}{TP + FP}$$

Table 2 Confusion matrix

	Reality		
Response system		Fault	No fault
	Alarm	True positive (TP)	False positive (FP)
	No alarm	Missed alarm (FN)	True negative (TN)

$$Accuracy = \frac{TP + TN}{all\ cases}$$

$$Sensitivy = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

3 Process and Alarm Data

The methods for alarm detection are used to improve the efficiency of the global process. However, a large amount of data must be provided to use these techniques.

There are two types of data that are fundamental to the management the alarm system:

- *Process data*: these are measurements of process variables at regular intervals, these are stored in a database and they provide information for the identification of the optimal alarm system.
- *Alarm data*: these are messages generated by the distributed control system (DCS), and they are stored in an alarm log.

This data is important to analyse, it can help to know the causes of the current alarm system overload. It is important to compare industrial data in a real environment with methods or techniques that are developed academically. For instance, Wang et al. explored the main factors behind this problem and they concluded that [19]: the chattering alarms frequently occur due to noise/disturbance, the alarm variables are incorrectly configured, the alarm design is isolated from related variables and the abnormality of the data is transmitted due to physical connections.

System performance and the alarm management lifecycle should be evaluated such as the runtime concept. Kondaveeti et al. [20] offers a tutorial to the alarms chatter, these are difficult to identify due to the poor design or incorrect configuration of the alarm method. A Chatter index is proposed to reduce the effort to identify and quantify chattering alarms. In reference [21], a quantitative measure is proposed to estimate the degree of chattering. The method for evaluating the chatter index is based on alarm parameters and statistical properties of the process variable. Process data is divided into approximate distribution characteristics, and each distribution is estimated separately. The distribution of process data is obtained by adding all run length distributions together. A mathematical function developed by analytical methods is intended to reduce chattering alarms.

Hu et al. proposed a framework for the combination of causality inference using process data and alarm data, and thus it helps to the operator to reduce the alarm flood [22]. Alarm data can be used to identify root alarm labels, and it reduces alarms that require attention. Root cause and effect analysis can be used to detect root cause

alarms. The random relationships can be detected by extracting the process variables associated with the root alarms. Finally, the root cause can be confirmed thanks to the causal map of the process variables and some knowledge of the process. The number of alarms is reduced with the method, since only root alarm tags are alerted. The operators can know the root cause quickly, because the causal relationship is detected. Process and alarm industrial data were applied, and the results presented good performances.

4 Alarm Flood

According to International Electrotechnical Commission (IEC, 2014) [23], an alarm flooding occurs when alarms appear on the control panels at a faster rate than the operator can manage them. It leads to determine the root cause of the alarm and the optimal control of the system.

A flood alarm is usually triggered by a primary event and its consequential events [24]. The root cause alarms should be distinguished from consequent alarms to reduce the number of alarms. The alarm data allows to make a list of the primary alarms and the process variables related to them. In addition, the causal relationships between the alarms are obtained with the alarm data. Subsequently, the process data will help to support or discern the root cause analysis.

The historical alarm data allows to use a new analysis method to eliminate alarm flooding [25]. These data are grouped according to a base of alarm occurrences. The alarm floods have similar patterns. If these patterns are analysed and classified, then this method can lead to the root cause of an anomaly. Therefore, the operator will have fewer false alarms and he will be able to react better to flooding alarms. Hu et al. applied a fast sequence alignment method to speed up the calculation and improve the computational efficiency of the algorithms [26]. The method is intended to be more sensitive to higher priority alarms, and it tends to ignore alarms that occur simultaneously to avoid flooding alarms. Through the set-based comparison is reduced unnecessary calculations by irrelevant alarm tags. The results obtained in industrial cases show that the method is faster than the existing algorithms and, therefore, the operators have more time to perform the correct operation and correct this failure.

An alarm that performs repeated transitions between the normal and the abnormal state is called a chattering alarm. This is mainly due to signal noise and because of the variable operates near the alarm limit. The chattering alarms cause many false alarms. It is proposed to redesign the control system, and that these alarms be eliminated by grouping. Consecutive alarms in a cluster are displayed spaced in a narrow time window, then become a single alarm. And only one alarm message will be sent to the operator for a single cluster when the alarm appears. This is a simple method to reduce alarm flooding.

Rodrigo et al. [27] are based on the previous line of work. They claim that by combining the alarm logging, analysing process data and connectivity, alarms can

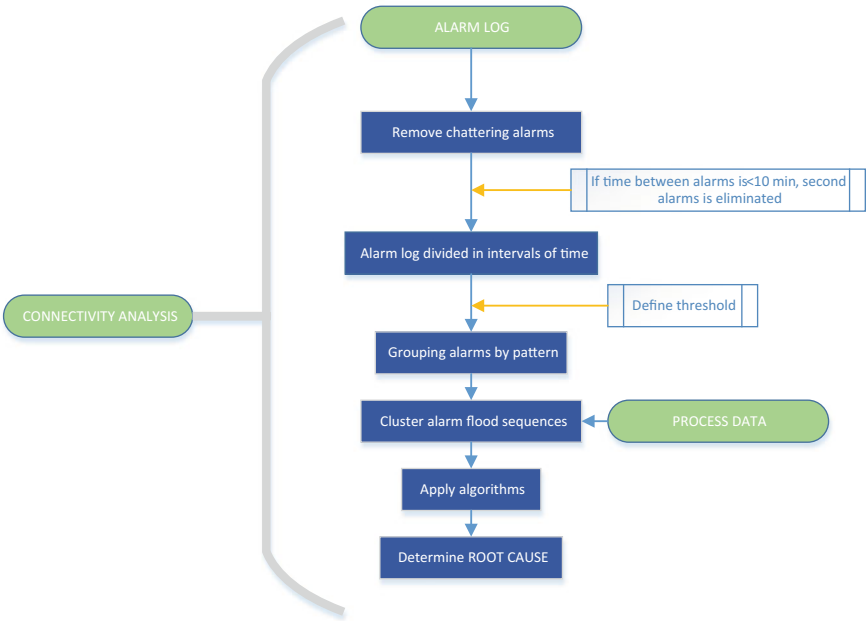


Fig. 1 Workflow for alarm systems

be grouped together, and their root alarm identified. Figure 1 shows the workflow to reduce the alarms flood.

The first step is remove chattering alarms, according to reference [25], the minimum permissible interval should be 10 min. If the elapsed time is shorter the second alarm is eliminated.

In the next step, the alarm log is divided in intervals of 10 min. An alarm threshold is set. It must be more than 10 alarms per time interval and per operator. Consecutive intervals are merged with more alarm occurrences than the defined threshold.

Using sequence pattern matching, the alarm flood sequences are grouped together. In this case, the method described in reference [28] is based on a modified Smith-Waterman (MSW) algorithm. Although, other algorithms can be applied, such as agglomerative hierarchical clustering (AHC).

The fourth step consists of grouping the flood alarm sequences, and a set of templates is created to cancel out the anomalies of all the clusters in the process.

Perhaps the last step is the most complicated, it should be noted that the causal alarm cannot be the first alarm, because when an alarm is triggered, it depends on the alarm setting limits. The time elapsed between the anomaly occurring and the alarm being triggered is probabilistic. Later, some algorithms are applied to determine the root cause of the alarm. There are many papers where different algorithms are applied [29, 30], the best algorithm will depend of the case study.

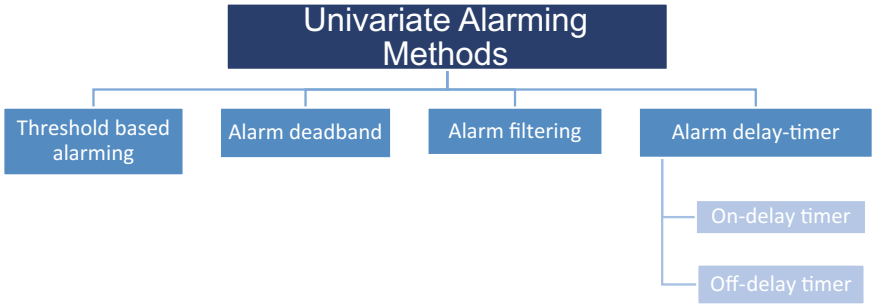


Fig. 2 Univariate alarming methods diagram

In summary, to reduce flooding alarm it is used: an alarm log, historical process data and connectivity analysis, to group the different alarms and determine the causal alarm.

There is no single solution to improve the alarm system, therefore, there are different workflows with various processes. For instance, there are signal-baseyvd methods, in this case the process variables are monitored and compared with thresholds (called alarm limits). They are currently the most widely used techniques in the industry and these are implemented in many modern distributed control systems (DCS).

There are also many classifications for alarm systems, some of the techniques applied are threshold design, data processing, multivariate process monitoring, model-based process monitoring, state-based priority setting [31–33].

Other classification of alarm systems depends on their design, that can be univariate and multivariate. Within the univariate design are: The alarm threshold; dead band; delay-timer, and; filtering (see Fig. 2). They are individually designed for each variable. In the multivariate design, alarms are combined linearly from various process variables.

Alarm flooding is difficult to suppress with delay timers or dead bands due to consequence alarms. Lai and Chen present an algorithm (extension of) for optimal alignment of multiple flood alarm sequences to obtain a common pattern of them [28, 34]. This new technique needs the following points: Similarity scoring functions; dynamic programming equation; tracking and alignment generation. They propose to develop new algorithms for combining online alarm messages with a database of patterns to alert operators in case of alarm flooding.

Data-driven method [35], concretely historical alarm data, is also employed to detect frequent patterns of alarm flooding. The results showed that the method is effective in finding patterns and reducing pattern redundancies. The holistic view of alarms is also employed for an intuitive understanding of alarm patterns.

The alarm flood sequence alignment (AFSA) methods provide fault inference from the assessment of the similarity of alarm sequences. Guo et al. proposed a new AFSA method, the match-based accelerated alignment (MAA), which analyses the

alarm coincidences [36]. It is important because its alignment results reveal to a large extent the real similarity of the alarm floods.

The alarm flood is a problem for the alarm system. There are several methods and techniques to avoid it, where the main ones are discussed in the following sections.

5 Long Standing Alarm

The long-standing alarms have several different definitions, for example ISA-18.2 defines them as "an alarm that remains in the alarm state for an extended period of time (e.g. 24 h)" [37]. According to EEMUA, 2013, an active alarm is considered a long duration alarm for a complete operating shift [38]. In general, the long-lasting alarm, as its name suggests, has a long alarm duration, but the authors do not agree on the thresholds for this time. In this chapter, three main causes of the generation of these alarms are indicated:

- Due to the modern computerized monitoring system, alarms are easily created by entering trigger point values, often implemented without special care and generate many misconfigured alarms.
- It is often not taken into account the start-up states, the average rate, etc., that have different demands and, therefore, different operating states, and are qualified as alarms when in fact they are not, e.g. when the equipment is switched off.
- The process variables experience variations in different states, but the alarm trigger points are constant. It would be interesting to compare the alarm variables with the measurements of the process variables and thus generate new alarm thresholds.

6 Graphical Methods

Alarm data display tools method are employed to detect the annoying alarms [39, 40], e.g. the High Density Alarm Plot (HDAP) and Alarm Similarity Color Map (ASCM). These graphical tools have proven their usefulness in identifying the chattering alarms.

HDAP presents the highest alarms for a given time. It is recommended to choose a sample size of 10 min, to follow the recommendations of the acceptable announcement rate according to EEMUA. This tool allows to emphasize through colour, for example red will show unacceptable chatter behaviour [41].

ASCM enables to be highlighted correctly, related and redundant alarms. This tool shows the alarms reorganized in terms of their similarity and time of occurrence. It depends on the time of analysis, number of higher alarms, type of union in the construction of the bunches and the method of arrangement of the leaves. This tool displays the data in a color-coded matrix and this allows the identification by groups of related alarms, which provide information on the interactions of the process.

Graphical representations provide valuable feedback to improve the alarm system and thus reduce false alarms. For example, Yang et al. used the pseudo data map according to [42]: (1) it is robust to false, missed and chattering alarms; (2) informs whether there is a positive or negative correlation and the similarity; (3) The pseudodata can be used in other statistical analyses to contrast the results obtained. The method consists of the following phases:

(a) The Gaussian kernel method is applied, and the binary alarm data generates continuous pseudo time series.
(b) A correlation colour map of pseudodata, or transformed data, is used for showing the set of correlated variables.
(c) Statistical methods are applied to find redundant alarm labels, or to group correlated alarms.

There are several difficulties to apply this method, such as parameter adjustment, the graph is sensitive, i.e. it requires some degree of freedom to optimize the display of the graph. However, it has been shown that this method is better than the alarm similarity colour map as long as the parameters are set properly.

7 Univariate Alarming Methods

The methods most commonly used are univariate alarming methods for alarm systems [43]. These methods are used because the information they show about a single signal is simple and clear, and operators can make decisions easily. However, for more complex alarms are needed other techniques such as multi-setpoint settings, mobile window, neural network method, etc. [44].

The most important univariate alarming methods are shown in Fig. 2

7.1 Alarm Filtering

The use of filters is widespread in real life because of they can be used for different proposes, for example: eliminating erroneous or undesirable data, reducing noise, extracting data characteristics, modifying the statistical distribution of data, grouping data according to their frequency. The most popular filters are the moving average, the exponentially moving average (EWMA) and the cumulative sum. Izadi et al. presented filters used to improve the receiver operating characteristic curve (ROC) [45].

Filtering techniques for alarm systems presents some disadvantages, e.g. measured by false alarm rate (FAR), missed alarm rate (MAR) and expected detection delay (EDD). Tan et al. [46] have worked with rank order filters to avoid the disadvantages. They have achieved two approaches when the PDF (probability density function) of raw data is known: performance curves of this filters can be calculated

directly and can be estimated the EDD, that is impossible for general filters. The experimental results have shown that the order of the filters offers a degree of freedom for the system design, and other if it is considered the size of the window. These results are limited to univariate alarms. Therefore, it is recommended to work with multivariate systems.

The accuracy is given by the false alarm rate, and the efficiency is related to the detection delay and the complexity of the methodology used [47]. Cheng et al. used a method to create an optimal filter design with the aim of improving performance [48]. The optimum performance curve leads in this case that the moving average filter is better than the linear filters. The authors propose as future work to study the performance of the generalized medium filter to obtain a robust optimal filter design method. Izadi et al. consider filtering, alarm delay or deadband to be simple techniques that can reduce annoying alarms and FAR [45].

7.2 Alarm Delay-Timer

Filters use a continuous function transformation, while alarm delay timers are the transformation of discrete functions. The timers are used for their simplicity and efficiency. They can reduce the FAR and MAR, but their disadvantage is that they suffer from a delayed response.

The main elements for univariate alarm design are: the set point; dynamic order, and; alarm algorithm. Su et al. proposed an alarm method with multiple setpoint delay timers [43]. This achieves a balance between accuracy and sensitivity of the alarm system by providing direct transitions from each delay timer sub-state to the alarm state. FAR, MAR and the averaged alarm delay (AAD) are reduced by this methodology. Xu et al. study the efficiency of a univariate system using FAR, MAR and AAD, with emphasis on the calculation of these rates [49]. The proposed method was applied to an industrial case, concluding that it can be used for power and petrochemical plants. Zang et al. employed an improved delay timer method, where the univariate alarm was configured with multiple commands and set points [50]. These timers had an alarm announcement set point and an alarm end set point over conventional alarm timers. Enhanced alarm timers have more design parameters, but present better performance according to the Markov chain. Markov chains are generally employed for random phenomena, being simple mathematical models. It applies to systems that are particularly dependent, as the state of the $n+1$ observation system depends only on the state of the system, i.e. changes in the system depend on the current state and not on the way it has been reached. Adnan et al. showed that the delay timers provide flexibility in the design of alarms [51]. The use of the delay timers is a common practice in the industry as it is a simple technique to reduce FAR, MAR and EDD.

Noise is one of the causes of chattering alarm. If a signal is well defined by its period and amplitude, but it contains noise and the noise is large enough to cross over the trigger point many times, then a chattering alarm occurs. Wang and Chen have

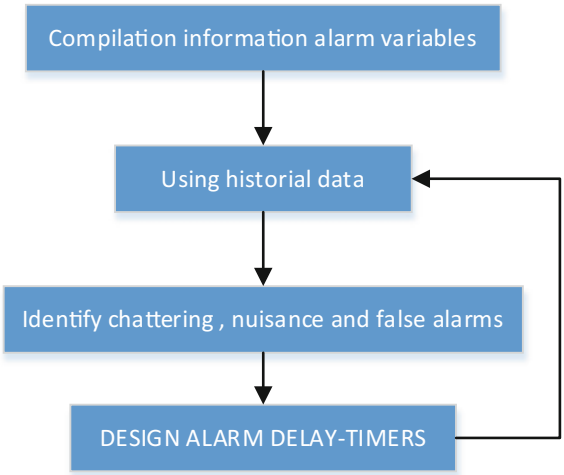


Fig. 3 Flowchart for the design and use the delay-timers

proposed an online method to detect and reduce chattering alarms due to oscillation [52]. The presence of oscillation can be determined through a revised chattering index and a method based on discrete cosine transform. Therefore, it is used an alarm setting or delay timer is used to reduce alarms. Wang and Chen [53] proposed a rule for detecting talking alarms caused by random noise, and other for repetitive alarms based on the duration and interval of alarms and by regular patterns. It uses the online method and the sample delay timer m to eliminate flicker and repeat alarms. The effectiveness of the method was tested using 3 industrial examples and according to FAR, MAR and AAD (Fig. 3).

8 Multivariate Alarming Methods

Some methods set the alarm limits by studying the correlations between the process data and the alarm data [54]. The multivariate statistical process control (MSPC) is a methodology that is applied for monitoring in many manufacturing processes [55]. It basically consists of three steps:

(1) The process is under normal operating conditions, historical data are collected and stored in the database, and a statistical model is developed.
(2) The control limits are fixed for the statistical model.
(3) If the online data exceeds the control limits, it will be qualified as a process failure.

Historical process data is subjected to multivariable statistical techniques to determine the control limits of the statistics of the study variables, if the actual values

exceed the control limit, then the point will be qualified as "out of control". This involves detection of faults, being the next step is to identify the root cause of the process fault [56].

False alarms can appear by different causes, where the failure of the alarm system and random effects are two of the main causes. System deficiency may be due to the difference between the statistical model and the real process. The random effects also may cause false alarms. There are some online-fluctuation being monitored in the process. They can cause actual variables to deviate from nominal values, and even though the process is working correctly, these false alarms can occur. Many authors have researched using a statistical approach to avoid randomly induced false alarms [57–60], e.g. Bernoulli, Binomial distributions, conventional method based on principal component analysis (PCA) [61], etc. However, the real variables of the process tend to be self-related, therefore, the approaches of modelling of time series are needed.

One of the main methods of multivariate analysis is the correlation method. In many processes, one variable can be affected by other variable or several variables, i.e. different alarm thresholds generate different alarm data and then different correlations. To optimize these multivariate alarm thresholds, numerous statistical methodologies or algorithms have been applied to demonstrate interactions between variables and determine correlated key variables for the optimization of alarm thresholds, grouped as:

- Grouping Variables
- Correlation Methods
- Advance Methods
- Intrusion Detection System (IDS).

9 Conclusions

An optimal alarm system should inform and guide, and each alarm should have a defined response and adequate time to allow the operator to respond adequately to that alarm. Alarms must be relevant, unique, prioritized and understandable. The alarm system must identify the alarm, sort it, set priorities and finally alert the operator if necessary, visually or audibly.

Due to the study of false alarms, it is concluded that three of the most important reasons for their existence are: (1) the process undergoes state changes such as switching on and off, this is set that abnormality and it propagates owing to physical connections; (2) the alarms are poorly configured and have redundant measurements, and; (3) exist causal relationships between the variables studied and alarm design is isolated from related variables.

There are many classifications on alarm systems, since depending on how they treat the information, the type of study variable, the algorithms applied, etc.

There are many types of alarm systems are used in the industry, however, false, annoying or chattering alarms have not yet been completely eliminated. Although many resources are devoted to this problem, an optimal solution has not yet been achieved. It will be possible to improve these methods by means of dynamic systems where the historical data provide feedback capable of handling the process correctly, due to the development of new technologies and the increase in data processing capacity.

Acknowledgements The work reported herewith has been financially supported by the Spanish Ministerio de Economía y Competitividad, under the Research Grants RTC-2016-5694-3 and DPI2015-67264-P.

References

1. ANSI. (2009). ISA-18.2-2009 management of alarm systems for the process industries. Durham, NC, USA: International Society of Automation.
2. Jiménez, A. A., Gómez Muñoz, C. Q., & García Márquez, F. P. (2018). Dirt and mud detection and diagnosis on a wind turbine blade employing guided waves and supervised learning classifiers. *Reliability Engineering and System Safety*.
3. Munoz, J. C., Márquez, F. G., & Papaelias, M. (2013). Railroad inspection based on ACFM employing a non-uniform b-spline approach. *Mechanical Systems and Signal Processing, 40,* 605–617.
4. EEMUA. (2007). *Alarm systems: A guide to design, management and procurement.* Engineering Equipment and Materials Users Association.
5. Hollifield, B. R., & Habibi, E. (2010). *Alarm management: A comprehensive guide: Practical and proven methods to optimize the performance of alarm management systems*. ISA.
6. Rothenberg, D. H. (2009). *Alarm management for process control: A best-practice guide for design, implementation, and use of industrial alarm systems*. Momentum Press.
7. Walker, B., Smith, K. D., & Kekich, M. D. (2003). Limiting shift-work fatigue in process control. *Chemical Engineering Progress, 99,* 54–57.
8. Gómez Muñoz, C. Q., Arcos Jimenez, A., García Marquez, F. P., Kogia, M., Cheng, L., Mohimi, A., & Papaelias, M. (2017). Cracks and welds detection approach in solar receiver tubes employing electromagnetic acoustic transducers. *Structural Health Monitoring*. https://doi.org/10.1177/1475921717734501.
9. Gómez Muñoz, C. Q., García Marquez, F. P., Lev, B., & Arcos, A. (2017). New pipe notch detection and location method for short distances employing ultrasonic guided waves. *Acta Acustica United with Acustica, 103,* 772–781.
10. de la Hermosa González, R. R., García Márquez, F. P., & Dimlaye, V. (2015). Maintenance management of wind turbines structures via MFCS and wavelet transforms. *Renewable and Sustainable Energy Reviews, 48*, 472-482.
11. Hu, W., Al-Dabbagh, A. W., Chen, T., & Shah, S. L. (2016). Process discovery of operator actions in response to univariate alarms. *IFAC-PapersOnLine, 49,* 1026–1031.
12. Severson, K., Chaiwatanodom, P., & Braatz, R. D. (2016). Perspectives on process monitoring of industrial systems. *Annual Reviews in Control, 42,* 190–200.
13. Marquez, F. G. (2006). An approach to remote condition monitoring systems management.
14. Arcos Jiménez, A., Gómez Muñoz, C. Q., & García Márquez, F. P. (2017). Machine learning for wind turbine blades maintenance management. *Energies*, *11*, 13.
15. García Márquez, F. P., Muñoz, G., Quiterio, C., Papelias, M., & Arcos Jiménez, A. (2015). A heuristic method for detecting and locating faults employing electromagnetic acoustic transducers.

16. Roberts, C., Márquez, F., & Tobias, A. (2010). A pragmatic approach to the condition monitoring of hydraulic level crossing barriers. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 224,* 605–610.
17. Izadi, I., Shah, S. L., Shook, D. S., & Chen, T. (2009). An introduction to alarm analysis and design. *IFAC Proceedings Volumes, 42,* 645–650.
18. Landgrebe, T. C., & Duin, R. P. (2008). Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30,* 810–822.
19. Wang, J., Yang, F., Chen, T., & Shah, S. L. (2016). An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems. *IEEE Transactions on Automation Science and Engineering, 13,* 1045–1061.
20. Kondaveeti, S. R., Izadi, I., Shah, S. L., Shook, D. S., Kadali, R., & Chen, T. (2013). Quantification of alarm chatter based on run length distributions. *Chemical Engineering Research and Design, 91,* 2550–2558.
21. Naghoosi, E., Izadi, I., & Chen, T. (2011). Estimation of alarm chattering. *Journal of Process Control, 21,* 1243–1249.
22. Hu, W., Chen, T., Shah, S. L., & Hollender, M. (2017). Cause and effect analysis for decision support in alarm floods. *IFAC-PapersOnLine, 50,* 13940–13945.
23. IEC. (2014). IEC 62682 management of alarm systems for the process industries. International Electrotechnical Commission (IEC).
24. Timms, C. (2009). Hazards equal trips or alarms or both. *Process Safety and Environmental Protection, 87,* 3–13.
25. Ahmed, K., Izadi, I., Chen, T., Joe, D., & Burton, T. (2013). Similarity analysis of industrial alarm flood data. *IEEE Transactions on Automation Science and Engineering, 10,* 452–457.
26. Hu, W., Wang, J., & Chen, T. (2015). Fast sequence alignment for comparing industrial alarm floods∗. *IFAC-PapersOnLine, 48,* 647–652.
27. Rodrigo, V., Chioua, M., Hagglund, T., & Hollender, M. (2016). Causal analysis for alarm flood reduction. *IFAC-PapersOnLine, 49,* 723–728.
28. Cheng, Y., Izadi, I., & Chen, T. (2013). Pattern matching of alarm flood sequences by a modified smith–waterman algorithm. *Chemical Engineering Research and Design, 91,* 1085–1094.
29. García Márquez, F. P., Chacón Muñoz, J. M., & Tobias, A. M. (2015). B-spline approach for failure detection and diagnosis on railway point mechanisms case study. *Quality Engineering, 27,* 177–185.
30. García Márquez, F. P., Pliego Marugán, A., Pinar Pérez, J. M., Hillmansen, S., & Papaelias, M. (2017). Optimal dynamic analysis of electrical/electronic components in wind turbines. *Energies, 10,* 1111.
31. García Márquez, F. P., Pedregal, D. J., & Roberts, C. (2015). New methods for the condition monitoring of level crossings. *International Journal of Systems Science, 46,* 878–884.
32. García Márquez, F. P., & Chacón Muñoz, J. M. (2012). A pattern recognition and data analysis method for maintenance management. *International Journal of Systems Science, 43,* 1014–1028.
33. de la Hermosa Gonzalez, R. R., García Márquez, F. P., Dimlaye, V., & Ruiz-Hernández, D. (2014). Pattern recognition by wavelet transforms using macro fibre composites transducers. *Mechanical Systems and Signal Processing, 48*, 339–350.
34. Lai, S., & Chen, T. (2017). A method for pattern mining in multiple alarm flood sequences. *Chemical Engineering Research and Design, 117,* 831–839.
35. Hu, W., Chen, T., & Shah, S. L. (2018). Detection of frequent alarm patterns in industrial alarm floods using itemset mining methods. *IEEE Transactions on Industrial Electronics*.
36. Guo, C., Hu, W., Lai, S., Yang, F., & Chen, T. (2017). An accelerated alignment method for analyzing time sequences of industrial alarm floods. *Journal of Process Control, 57,* 102–115.
37. Stauffer, T., Sands, N., & Dunn, D. (2010). Alarm management and ISA-18–a journey, not a destination. In: Texas A&M Instrumentation Symposium.
38. Ávila, S., & Pessoa, F. (2015). Proposition of review in EEMUA 201 and ISO standard 11064 based on cultural aspects in labor team, lng case. *Procedia Manufacturing, 3,* 6101–6108.

39. Kondaveeti, S. R., Izadi, I., Shah, S. L., Black, T., & Chen, T. (2012). Graphical tools for routine assessment of industrial alarm systems. *Computers and Chemical Engineering, 46,* 39–47.
40. Kondaveeti, S. R., Izadi, I., Shah, S. L., & Black, T. (2010). Graphical representation of industrial alarm data. *IFAC Proceedings Volumes, 43,* 181–186.
41. EEMUA. (1999). Alarm systems: A guide to design, management and procurement. Engineering Equipment and Materials Users Association London.
42. Yang, F., Shah, S. L., Xiao, D., & Chen, T. (2012). Improved correlation analysis and visualization of industrial alarm data. *ISA Transactions, 51,* 499–506.
43. Su, J., Guo, C., Zang, H., Yang, F., Huang, D., Gao, X., et al. (2018). A multi-setpoint delay-timer alarming strategy for industrial alarm monitoring. *Journal of Loss Prevention in the Process Industries, 54,* 1–9.
44. Jiménez, A. A., Gómez Muñoz, C. Q., García Marquez, F. P., & Zhang, L. (2017). Artificial intelligence for concentrated solar plant maintenance management. In: *Proceedings of the Tenth International Conference on Management Science and Engineering Management*, pp. 125–134. Springer.
45. Izadi, I., Shah, S. L., Shook, D. S., Kondaveeti, S. R., & Chen, T. (2009). A framework for optimal design of alarm systems. *IFAC Proceedings Volumes, 42,* 651–656.
46. Tan, W., Sun, Y., Azad, I. I., & Chen, T. (2017). Design of univariate alarm systems via rank order filters. *Control Engineering Practice, 59,* 55–63.
47. García Márquez, F. P. (2010). A new method for maintenance management employing principal component analysis. *Structural Durability and Health Monitoring, 6,* 89–99.
48. Cheng, Y., Izadi, I., & Chen, T. (2013). Optimal alarm signal processing: Filter design and performance analysis. *IEEE Transactions on Automation Science and Engineering, 10,* 446–451.
49. Xu, J., Wang, J., Izadi, I., & Chen, T. (2012). Performance assessment and design for univariate alarm systems based on FAR, MAR, and AAD. *IEEE Transactions on Automation Science and Engineering, 9,* 296–307.
50. Zang, H., Yang, F., & Huang, D. (2015). Design and analysis of improved alarm delay-timers. *IFAC-PapersOnLine, 48,* 669–674.
51. Adnan, N. A., Cheng, Y., Izadi, I., & Chen, T. (2013). Study of generalized delay-timers in alarm configuration. *Journal of Process Control, 23,* 382–395.
52. Wang, J., & Chen, T. (2013). An online method for detection and reduction of chattering alarms due to oscillation. *Computers and Chemical Engineering, 54,* 140–150.
53. Wang, J., & Chen, T. (2014). An online method to remove chattering and repeating alarms based on alarm durations and intervals. *Computers and Chemical Engineering, 67,* 43–52.
54. Pliego Marugán, A., García Márquez, F. P., & Lev, B. (2017). Optimal decision-making via binary decision diagrams for investments under a risky environment. *International Journal of Production Research, 55,* 5271–5286.
55. Peres, F. A. P., & Fogliatto, F. S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers and Industrial Engineering, 115,* 603–619.
56. Gómez Muñoz, C. Q., García Márquez, F. P., & Sánchez Tomás, J. M. (2016). Ice detection using thermal infrared radiometry on wind turbine blades. *Measurement, 93,* 157–163.
57. Abraham, B., & Chuang, A. (1993). Expectation-maximization algorithms and the estimation of time series models in the presence of outliers. *Journal of Time Series Analysis, 14,* 221–234.
58. Chen, T., & Sun, Y. (2009). Probabilistic contribution analysis for statistical process monitoring: A missing variable approach. *Control Engineering Practice, 17,* 469–477.
59. Singhal, A., & Seborg, D. E. (2000). Dynamic data rectification using the expectation maximization algorithm. *AIChE Journal, 46,* 1556–1565.
60. Chen, T. (2010). On reducing false alarms in multivariate statistical process control. *Chemical Engineering Research and Design, 88,* 430–436.
61. García Márquez, F. P., & García-Pardo, I. P. (2010). Principal component analysis applied to filtered signals for maintenance management. *Quality and Reliability Engineering International, 26,* 523–527.