Exploring applications of Big Data Analytics in Supply Chain Management

Truong Van Nguyen

The thesis submitted in partial fulfilment of the requirements of the University of Greenwich to the degree of Doctor of Philosophy

Business school

University of Greenwich, London, United Kingdom

September 2019

Declaration

I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me, and is not concurrently being submitted for any degree other than that of Doctor of Philosophy being studied at the University of Greenwich. I also declare that this work is the result of my own investigations, except where otherwise identified by references and that the contents are not the outcome of any form of research misconduct.

Candidate: Truong Van Nguyen

First Supervisor: Dr. Li Zhou

Second Supervisor: Dr. Yong Lin

Abstract

Empirical evidence demonstrates many benefits of Big data analytics (BDA) in supply chain management (SCM), including reduced operational costs, improved SC agility, and increased customer satisfaction. However, reports show that the BDA adoption of companies in SCM is relatively low, and the main reason for this is lack of understanding of how it can be implemented to address specific business problems.

Therefore, the aim of this thesis is to explore new applications of BDA to support the data-driven decision making in SCM. Particularly, the thesis addresses four research objectives: (1) to conduct a literature review that summarises what and how BDA has been applied within the SCM context. As a result, several research gaps are revealed, which leads to future research directions; (2) to develop a comprehensible, data-driven demand prediction of remanufactured products. Validated with a real-world Amazon dataset, the result shows that the proposed approach can produce a highly accurate and robust prediction of product demand, as well as providing insights into the non-linear effect of online market factors on demand; (3) to develop a prescriptive price optimisation model by extending the proposed demand prediction with a mixed integer linear programme to optimise promotional pricing decisions. The result shows that the obtained optimal promotional price solution could increase both sales and revenue; (4) Finally, the thesis proposes a data-driven prescriptive approach for large-scale optimisation problems, based on the hybrid approach combining association rule mining and complex network theory. For validation, the proposed model is applied to optimise the large-scale dry port location problems in Mainland China in the context of the Belt and Road Initiatives (BRI). The dry port solution obtained from the model is realistic and applicable as it accurately pinpoints key locations in the real BRI development plans.

The contribution of this thesis is multifaceted. Theoretically, the thesis serves as a good starting point for researchers to build up the foundation of BDA, which enables to develop a machine learning-based approach to tackle the established research problems. Practically, the thesis facilitates the data-driven decision making across industries such as online marketing strategy development for remanufactured products, daily promotional planning for retailers, and logistics network design for dry port planners.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Dr. Li Zhou for the continuous support of my PhD and academic career. I'm very grateful for her patience, inspiration and immense knowledge. Her guidance has helped me in all stages of my research. I honestly could not have asked for a better PhD mentor.

My sincere thanks also goes to Prof. Petros Ieromonachou, Prof. Alain Chong, Dr. Yong Lin, Dr. Virginia Spiegler, and Dr. Meng Meng. Without their precious support and encouragement, I would not be able to conduct this research.

Last but not least, my deepest gratitude to my family, especially my parents, for their continuous love and support. I am forever indebted to them for giving me the best opportunities that they have and selflessly encouraging me to seek whatever makes me happy.

Table of Contents

Declaration ... ii

Abstract ... iii

Acknowledgements ... iv

Table of Contents ... v

List of Figures ... viii

List of Tables ... x

Chapter 1 Introduction ... 1

1.1 Research background and motivation ... 1

1.2 Research aim and objective ... 2

1.3 Thesis structure ... 2

1.4 List of publications during the PhD study period ... 4

1.4.1 Papers published in academic journals ... 4

1.4.2 Papers published in academic conferences ... 5

Chapter 2 Literature Review ... 6

2.1 Introduction ... 6

2.2 Literature review methodology ... 7

2.2.1 Material collection ... 8

2.2.2 Descriptive analysis ... 9

2.2.3 Category selection ... 11

2.3 Material evaluation ... 13

2.3.1 Reviewing by SC functions ... 13

2.3.2 Reviewing by level of data analytics ... 16

2.3.3 Reviewing by types of BDA models ... 17

2.3.4 Reviewing by BDA techniques ... 19

2.4 In-depth content analysis based on each SC function ... 20

2.4.1 Procurement ... 20

2.4.2 Manufacturing ... 21

2.4.3 Logistics and Transportation ... 24

2.4.4 Warehousing ... 26

2.4.5 Demand management ... 28

2.4.6 General SCM ... 30

2.5 Discussion ... 31

2.6 Future directions ... 35
2.7 Updated literature review after 2017 until 2019 ... 38
2.8 Summary ... 40
Chapter 3 Research Methodology ... 42
3.1 Fundamentals of research paradigm ... 42
3.1.1 Ontology ... 42
3.1.2 Epistemology ... 43
3.1.3 Methodology ... 44
3.2 Data-driven science – a new paradigm in Big Data era ... 46
3.3 Summary ... 47
Chapter 4 Data-driven demand prediction of remanufactured products ... 48
4.1 Introduction ... 49
4.1.1 Research background and motivation ... 49
4.1.2 Review on market development of remanufactured products ... 50
4.1.3 Review on ML approaches for regression problems ... 52
4.1.4 Research objective and contribution ... 53
4.2 Methodology ... 54
4.3 Data collection and use of variables ... 55
4.3.1 Research context and data collection ... 55
4.3.2 Theoretical background and variable description ... 56
4.4 Data preparation ... 66
4.4.1 Structured data preparation ... 66
4.4.2 Unstructured textual data preparation ... 67
4.5 Model development and validation ... 68
4.5.1 Variable selection ... 68
4.5.2 Regression tree predictive model descriptions ... 69
4.5.3 Hyperparameter tuning ... 72
4.5.4 K-fold cross validation (CV) ... 72
4.6 Model evaluation and deployment ... 73
4.6.1 Model evaluation ... 73
4.6.2 Model deployment ... 73
4.7 Results and discussions ... 76
4.7.1 Result of variable selection step ... 77
4.7.2 Model predictive performance evaluation ... 80
4.7.3 Model deployment – interpretation of results and managerial insights ... 84

4.8 Robustness check 93
4.8.1 Test 1: Using the number of customer reviews 93
4.8.2 Test 2: Using the new dataset in different time periods 95
4.8.3 Test 3: Using different types of remanufactured products 95
4.9 Summary 98
Chapter 5 Data-driven prescriptive optimisation of promotional pricing 99
5.1 Introduction 99
5.2 Methodology 102
5.3 Experimental result 106
5.3.1 Data understanding 106
5.3.2 Data preparation 107
5.3.3 Result from the first stage – Sales forecasting 110
5.3.4 Result from the second stage – promotional price optimisation 112
5.4 Summary 115
Chapter 6 Data-driven prescriptive optimisation of large-scale dry port network design 117
6.1 Introduction 118
6.2 Review of dry port location studies 122
6.3 Methodology 124
6.3.1 Constructing international transaction database 125
6.3.2 Stage 1: ARMEC algorithm 126
6.3.3 Stage 2: GCS algorithm 130
6.4 Experiment 134
6.4.1 Case study: Dry port developments under China’s Belt and Road initiative ..134
6.4.2 Data collection 136
6.4.3 Experiment results and discussion 138
6.5 Robustness check 146
6.6 Summary 149
Chapter 7 Conclusion 150
7.1 Summary of findings 150
7.2 Contributions 152
7.3 Limitations and future directions 153
References 156
Appendix 192

List of Figures

Figure 1.1 - Thesis structure 4
Figure 2.1 - Systematic literature search process....................9
Figure 2.2 - Distribution of reference papers by year....................10
Figure 2.3 - Distribution of reference papers by publication....................10
Figure 2.4 - Classification framework 12
Figure 2.5 - Distribution of each SC function in BDA....................13
Figure 2.6 - Distribution of analytics level by year 16
Figure 3.1 – Epistemologies with positivist and interpretivist influence 44
Figure 4.1 - Methodology framework....................54
Figure 4.2 - RMSE reduction when adding and/or removing predictors using RF 79
Figure 4.3 - Variable importance ranked by the reduction in Z-score (y-axis) of each independent variable (x-axis) in the Boruta algorithm 80
Figure 4.4 – Prediction results in the training (boxplots) and testing (red points) stage with the ten-fold cross validation....................82
Figure 4.5 - VIR result by using IFSA method....................85
Figure 4.6 - The PDP of number of service failures and service successes....................86
Figure 4.7 - The PDP of overall ratings and customer review sentiments 89
Figure 4.8 - The PDP of number of product description positivity 91
Figure 4.9 - The PDP of number of answered questions, number of helpful votes and band equity....................92
Figure 4.10 – Result of robustness test 1 94
Figure 4.11 – Result of robustness test 2 96
Figure 5.1 – Methodology framework 102
Figure 5.2- Distribution of sales transaction records by retailer country 109
Figure 5.3 - Prediction results in the training (boxplots)....................112
Figure 5.4 – Sales and Revenues corresponding to the optimal prices....................114
Figure 6.1 - The ARMEC-GCS approach....................125
Figure 6.2 - Example of international purchasing network used in this study 129
Figure 6.3 - The Belt and Road framework 135
Figure 6.4 - Overall description of the Alibaba international transaction database....................137

Figure 6.5 - An international purchasing network that visualises 3,110 association rules (red nodes) and their associated Chinese supplying cities (green nodes). 140

Figure 6.6 - Dry port locations and their coverage areas by the ARMEC-GCS approach 142

Figure 6.7 - Dry port locations based on the results of the GCS-only approach. 146

List of Tables

Table 2.1 - Literature review classification 11
Table 2.2 - Summary of literature by SC functions 15
Table 2.3 - Level of analytics in each SC function 17
Table 2.4 - Distribution of articles by BDA models 18
Table 2.5 - Distribution of articles by BDA techniques 19
Table 2.6 - Summary of literature on BDA in procurement 21
Table 2.7 - Summary of literature on BDA in manufacturing 23
Table 2.8 - Summary of literature on BDA in Logistics/Transportation 24
Table 2.9 - Summary of literature on BDA in warehousing 28
Table 2.10 - Summary of literature on BDA in demand management 29
Table 2.11 - Summary of literature on BDA in “General SCM” papers 30
Table 2.12 - Summary of updated literature between 2018-2019 39
Table 4.1 - Summary of studies on market development of remanufactured products 50
Table 4.2 - Examples of the positivity of product description 59
Table 4.3 - Statistic description of data 77
Table 4.4 - RFE ranking of variable relevance based on predictive performance 78
Table 4.5 - Model hyperparameter tuning based on ten-fold cross-validation 81
Table 4.6 - The t-test result for MAE 83
Table 4.7 - The t-test result for RMSE 84
Table 4.8 - The t-test result for R^2 84
Table 4.9 - Summary of main findings in this research 97
Table 5.1 - Statistic description of data 107
Table 5.2 - Summary of product lines and product types by transaction counts 108
Table 5.3 - The RFE ranking of variable relevance based on predictive performance 110
Table 5.4 - Model hyperparameter tuning based on ten-fold cross-validation 111
Table 5.5 - Promotion cap value 113
Table 5.6 - Optimal price-promotion plan based on case 2 115
Table 6.1 - Summary of literature on dry port location problem 123
Table 6.2 - Example of the international transactional database 126
Table 6.3 - Statistical description of the distribution for 3,110 rules 139

Table 6.4 - Top 10 out of 3,110 rules sorted by confidence 139

Table 6.5 - Dry port locations with assigned communities using the ARMEC-GCS 143

Table 6.6 - Major international gateway ports under the BRI and suggested partnerships with dry ports based on the results of this experiment...145

Table 6.7 - Dry port locations with assigned communities using the GCS-only approach ...147

Table 6.8 - Sensitivity analysis result 148

Chapter 1 Introduction

1.1 Research background and motivation

With the tremendous development of information and communication technologies (ICTs), almost every sector of business has been affected by the abundance of information, which is called big data (BD). One of the most common ways to define BD is through the 5Vs: volume, variety, velocity, veracity, and value (Wamba et al., 2015; Assunção et al., 2015; Emani et al., 2015). *Volume* refers to the magnitude of data, which has exponentially increased, remaining a challenge to the capacity of existing storage devices (Chen and Zhang, 2014). *Variety* refers to the fact that data can be generated from heterogeneous sources (e.g., sensors, the Internet of things (IoT), mobile devices, online social networks, etc.) in structured, semi-structured, and unstructured formats (Tan et al., 2015). *Velocity* refers to the speed of data generation and delivery, which can be processed in batch, real-time, nearly real- time, or streams (Assunção et al., 2015). *Veracity* stresses the importance of data quality and level of trust due to the concern that many data sources (e.g., social network sites) inherently contain a certain degree of uncertainty and unreliability (Gandomi and Haider, 2015; IBM, 2012; White, 2012). Finally, *Value* refers the process of revealing underexploited values from BD (IDC, 2012; Oracle, 2012).

Among those 5Vs, veracity and value, which represent the rigorousness of Big Data Analytics (BDA), are particularly important because without data analysis, other BD processing aspects such as collection, storage and management would not create much value (Huang et al., 2015; Chen and Zhang, 2014; Babiceanu and Seker, 2016).

BDA refers to the use of advanced analytics techniques to extract valuable knowledge from vast amounts of data, facilitating data-driven decision making (Tsai et al., 2015). Supply chain management (SCM) has been extensively applying a large variety of technologies, such as sensors, bar codes, radio frequency identification (RFID), the IoT, etc. to integrate and coordinate every linkage of the chain. Therefore, not surprisingly, it is one of the business areas that has been drastically revolutionized by BDA. Empirical evidence demonstrates multiple advantages of BDA in SCM, including reduced operational costs, improved SC agility, and increased customer satisfaction (Sheffi, 2015). Although the expectation of BDA adoption to enhance SC performance is rather high, a recent report found that only 17% of enterprises had implemented BDA in one or more supply chain (SC) functions (Wang et al., 2016). One of the main reasons for the low uptake of BDA is a lack of understanding of how it can be implemented to address specific business problems. This *motivates* the author to explore new applications of BDA in SCM.

1.2 Research aim and objective

To stimulate the practical use of BDA-enabled SCM, the aim of this thesis is to develop methodologies that can explore what and how BDA can be used to support data-driven decision making in SCM. In particular, the thesis pursues four research objectives:

(1) Conduct a literature review on the applications of BDA in SCM.
(2) Develop a comprehensible, data-driven sales prediction model.
(3) Develop a prescriptive, data-driven optimisation model of promotional pricing.
(4) Develop a prescriptive, data-driven optimisation model of large-scale dry port network design.

1.3 Thesis structure

A brief overview of the thesis structure and how each chapter connects to the research is described as below.

Chapter 1: Introduces the research background and motivation of the thesis relating to the practical use of BDA in SCM. The research aim and objectives are then formulated.

Chapter 2: Provides a literature review that summarise the current applications and the used methodologies of BDA in SCM, thus revealing current research gaps and pointing out how BDA can be further explored in future research.

Chapter 3: Outlines the research methodology used to carry out this research. It briefly describes fundamental of research paradigm and data-driven science.

Chapter 4: Proposes a comprehensible machine learning-based approach that can (1) produce an accurate and robust demand prediction of remanufactured products, and (2) open the "black box" in the machine learning-based prediction to unveil the unknown, non-linear market behaviour of remanufactured products. The model is applied to Amazon case study for validation. One of the research gaps identified from the literature review in *Chapter 2* that the application of BDA on reverse logistics and closed-loop supply chain (CLSC) is largely underexploited. Therefore, the proposed demand prediction model here is among the pioneering studies that promote the adoption of BDA into CLSC area.

Chapter 5: Proposed a prescriptive analytics empowered by integrating predictive analytics in *Chapter 4* with the mixed integer linear programming (MILP) optimisation model. For validation, the approach is applied into the promotional price optimisation in a real case study.

Chapter 6: Proposed a prescriptive analytics empowered by integrating descriptive analysis based on association rule mining with modularity-based, heuristic algorithm in complex network theory to optimise the large-scale dry port location problem. This approach overcomes the issue of scalability limitation of the prescriptive model in *Chapter 5*.

Chapter 7: provides the overall conclusion of the thesis, including a summary of the main findings, theoretical and practical contributions, limitations, and future research directions.

The structure of the thesis is illustrated in Figure 1.1.

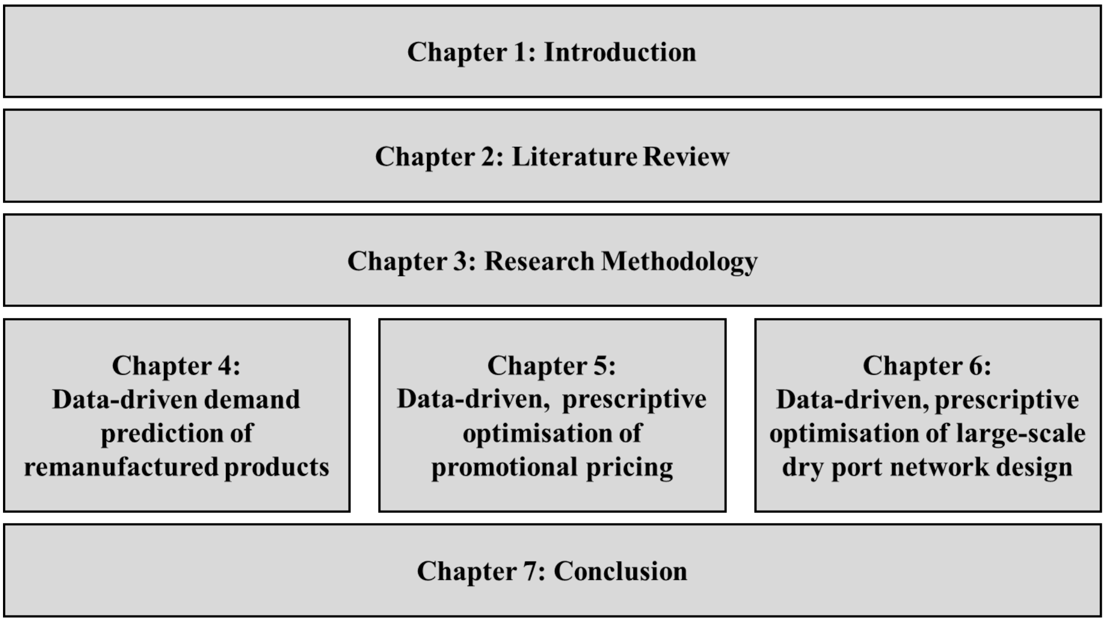

Chapter 1: Introduction
Chapter 2: Literature Review
Chapter 3: Research Methodology
Chapter 4: Data-driven demand prediction of remanufactured products
Chapter 5: Data-driven, prescriptive optimisation of promotional pricing
Chapter 6: Data-driven, prescriptive optimisation of large-scale dry port network design
Chapter 7: Conclusion

Figure 1.1 - Thesis structure

1.4 List of publications during the PhD study period

This section provides a list of publications during the author's PhD study period. They are categorised as journal papers and conference papers.

1.4.1Papers published in academic journals

Nguyen, T. V., Zhou, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2017). Big data analytics in supply chain management: A state-of-the-art literature review. *Computers & Operations Research*, 98, pp.254-264. https://doi.org/10.1016/j.cor.2017.07.004

Nguyen, T. V., Zhou, L., Chong, A., Li, B. and Pu, X. (2019). Predicting Customer Demand for Remanufactured Products: A Data-Mining Approach. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2019.08.015

Hu, Y., Zhou, L., Xie, C., Wang, G-J. and **Nguyen, T. V.** (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, 211, pp. 22 - 33. https://doi.org/10.1016/j.ijpe.2019.01.032

Nguyen, T. V., Zhang, J., Zhou, L., Meng, M. and He, Y. (2019). A data-driven optimization of large-scale dry port locations using the hybrid approach of data mining and complex network theory. *Transportation Research Part E: Logistics and Transportation Review*. (Under-reviewed with minor changes).

1.4.2 Papers published in academic conferences

Nguyen, T. V., L. Zhou. and Y. Lin. (2017) 'A multi-objective, multi-product and multi-transportation mode sustainable closed-loop supply chain network design'. *2016 International Conference on Logistics, Informatics and Service Sciences (LISS).* Sydney, NSW, Australia. 16 - 27 April. IEEE. pp. 1 - 6.

Chapter 2 Literature Review

As suggested in the previous chapter, the low uptake of BDA in practice is largely due to a lack of understanding about what and how BDA can be applied. Hence, the chapter provides a literature review to summarise the current applications and the used methodologies of BDA, thus pointing out how BDA can be further explored in future research.[1]

In particularly, Section 2.1 introduces the importance of the review and sets out the research aim and research questions. Section 2.2 describes the review methodology used for the literature search and delimitation, as well as proposing the classification framework of the literature. Section 2.3 is the material evaluation in which the literature is categorised and analysed based on the proposed classification framework. Section 2.4 provides the in-depth content analysis of the literature. Section 2.5 discusses the findings and Section 2.6 suggests the directions for future research. Section 2.7 includes the updated literature review between 2018 and 2019. Section 2.8 is the conclusion with the limitation of this literature review.

2.1 Introduction

There are a number of literature review papers about BDA in SCM. However, most of them focus on either a specific operational function in SCM or purely from technical aspects. For example, O'Donovan et al. (2015) examined BDA adoptions in manufacturing by using a systematic mapping review methodology. Wamba et al. (2015) developed a framework for applying BDA, particularly in humanitarian logistics, through a literature review. They then used a case study of the Australian State Emergency Service to verify the framework. A similar review approach was also utilized by Dutta and Bose (2015) in a manufacturing case study, and Olson (2015) reviewed data mining techniques for SCM. However, to the best of the

[1] Chapter 2 has been published during the first year of the author's PhD. See the reference:

Nguyen, T. V., Zhou, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2017). Big data analytics in supply chain management: A state-of-the-art literature review. *Computers & Operations Research*, 98, pp.254-264.

knowledge, literature reviews which takes a holistic perspective of SCM as a whole and cross-maps with BDA techniques is rather scarce (Olson, 2015; Addo-Tenkorang and Helo, 2016; Hazen et al., 2016; Wang et al., 2016). Thus, this literature review aims to provide a overview of where and how BDA has been applied in SCM, by mapping BDA models and techniques to SC functions. More specifically, the literature review attempts to address the following four research questions:

(1) What areas in SCM that BDA is being applied?

(2) Which level of analytics is BDA implemented in these SCM areas?

(3) What types of BDA models are used?

(4) What BDA techniques are employed to develop these models?

The next chapter describe the review methodology used in this study.

2.2 Literature review methodology

To address the aforementioned research questions, the review methodology is based on the content analysis approach proposed by Mayring (2008). This approach has been adopted by a number of highly cited review papers in SCM literature, such as Seuring and Muller (2008), Seuring (2013) and Govindan et al. (2015). In particular, the review is systemically conducted in accordance with the four-step iterative process (Seuring, 2013). The first step is material collection (Section 2.2.1), which involves defining and delimiting rules about the document types and unit of analysis to be selected. The second step uses descriptive analysis (Section 2.2.2) to assess formal aspects of the material (e.g., time published and publication). The third step is category selection (Section 2.2.3), which aims to construct a classification framework with structural dimensions and analytic categories. Each structural dimension includes a range of analytic categories, and each analytic category is a key issue or topic that arose from the literature analysis. The final step is material evaluation (presented in Section 2.3) in which the content of the material is analysed and sorted in accordance with the proposed classification framework in order to identify significant findings and enable their interpretation.

2.2.1Material collection

Before searching for articles, it is essential to identify an effective set of keywords that can capture the synthesis of the existing literature related to the research topic. The author classified keywords into two groups:

- Group 1: words related to BDA: “Big Data”, “Data analytics”, “Data mining”, “Machine learning”, “Descriptive analytics”, “Predictive analytics”, and “Prescriptive analytics”.
- Group 2: words related to SCM: “Supply Chain”, “Purchasing”, “Procurement”, “Manufacturing”, “Inventory”, “Storage assignment”, “Order picking”, “Logistics”, “Transport”, and “Marketing”.

Note that “inventory”, “storage assignment,” and “order picking” are three major functions of warehousing operations in SC. The reason the author used these specific terms rather than “warehouse” was to avoid confusion with “data warehousing,” a well-established technical BDA research topic.

The search was conducted based on all possible pairs between the two types of keywords. The keyword strings were searched in titles, abstracts, and keywords within the timeline from 2010 to 2017 on well-known academic databases, i.e., Science Direct, Emerald, Scopus, EBSCO, and IEEE Xplore. The author choses this particular timeline because although the term “Big Data” emerged in 2007, BDA’s initial transition into a global phenomenon actually began in 2010, according to a report from McKinsey Global Institute (Manyika et al., 2011).

The initial search generated a total of 1,550 papers. After eliminating duplicated results, the total number of papers dropped to 865 papers. Then, the author checked the overall relevance of the remaining papers by removing those that did not contain keywords related to both BDA and SCM functions in the title or abstract. This screening process reduced the number of papers to 594. The remaining papers were then filtered based on inclusion and exclusion criteria. These criteria were developed and justified by the authors in order to minimize the impact of the subjective bias, as suggested by Tranfield et al., (2003). 408 papers met the inclusion criteria and went into the final filtering with exclusion criteria.

After critically reading the introduction and discussion section of the remaining 408 papers, the following exclusion rule was applied: removing the papers that only mention the application of BDA on SCM as a fleeting point of reference or as collateral research topics. In fact, many BDA-related papers only point out the potential benefits and applications of BDA to SCM without investigating *how* they are actually implemented, i.e extracting, loading and transferring massive datasets, and use advanced analytic techniques to support SCM decision-making. In the end, 86 papers were kept for full review. Figure 2.1 summarizes the systematic articles search and selection process.

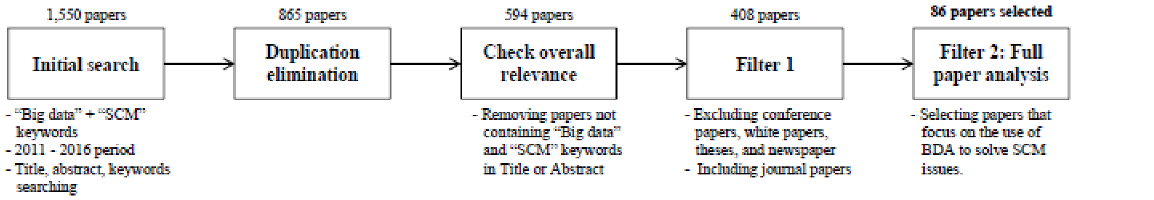

1,550 papers
Initial search
- "Big data" + "SCM" keywords
- 2011 - 2016 period
- Title, abstract, keywords searching
865 papers
Duplication elimination
594 papers
Check overall relevance
- Removing papers not containing "Big data" and "SCM" keywords in Title or Abstract
408 papers
Filter 1
- Excluding conference papers, white papers, theses, and newspaper
- Including journal papers
86 papers selected
Filter 2: Full paper analysis
- Selecting papers that focus on the use of BDA to solve SCM issues.

Figure 2.1 - Systematic literature search process

2.2.2 Descriptive analysis

This section examines the distribution of the reference papers by the time period and by publication. Figure 2.2 indicates that the number of papers has increased over the last five years, and especially upsurged since 2014. This observation is consistent with the frequency distribution of documents containing "big data" found in Gandomi and Haider (2015). It suggests that the application of BDA in SCM is a fast-growing and fruitful research field which has been promoted by several recent special issue calls.

In term of publication, the selected 86 papers are from 44 different journals in which only 14 published more than one paper. Figure 2.3 illustrates the distribution of the reference papers in these 14 journals. As can be seen from the figure, the papers published on the BD–SCM topics are spread out across a great variety of journals. A large number of papers focus on three fundamental SC functions; i.e., marketing, transportation, and manufacturing, indicating the relevance and importance of BDA applications across the SC. Furthermore, this significance has attracted real interest from highly regarded academics because most of these papers were published by journals with high impact factors.

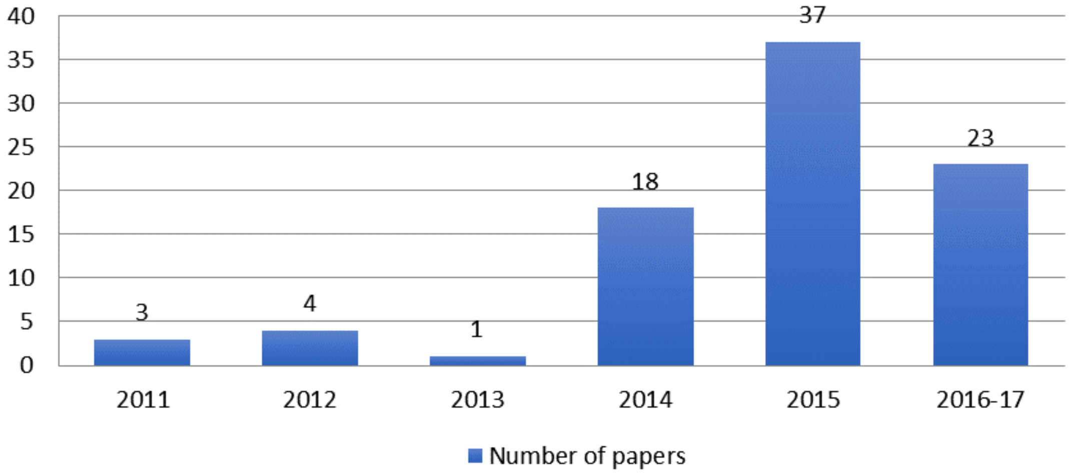


Figure 2.2 - Distribution of reference papers by year

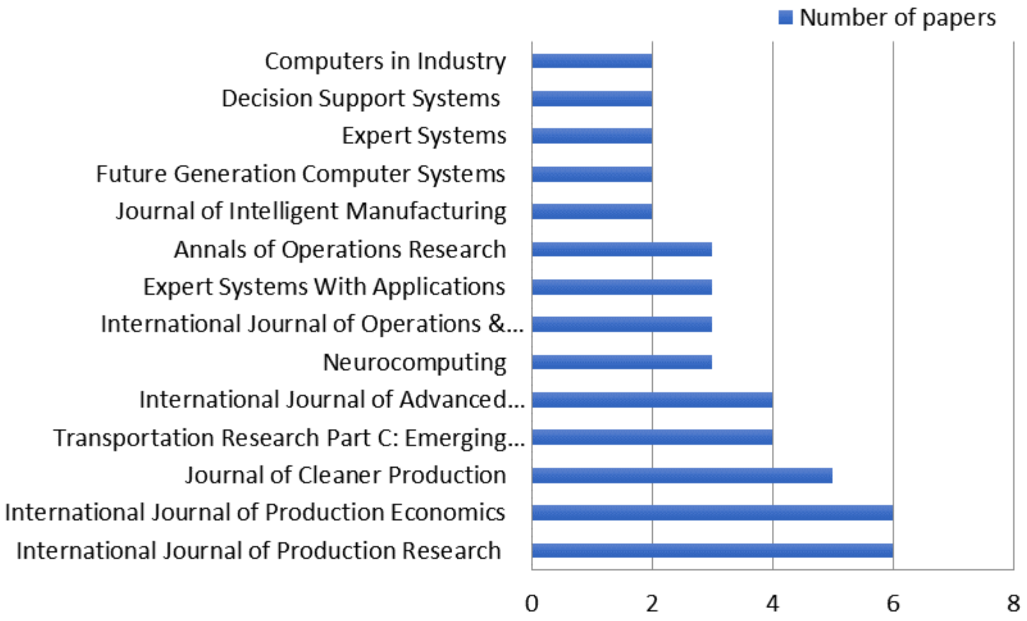


Figure 2.3 - Distribution of reference papers by publication

2.2.3 Category selection

The category selection step entailed conceptualizing the classification framework, which is comprised of structural dimensions and analytic categories. In order to address the proposed research questions, the author selects four structure dimensions to layer the classification framework, namely *SCM functions, levels of data analytics, BDA models,* and *BDA techniques*. Each structure dimension consists of analytic categories which refer to key topics derived deductively based on various classification frameworks from prior literature. Table 2.1 summarises these analytic categories.

Table 2.1 - Literature review classification

Structure dimension		**Analytic categories**
SC functions	**Procurement**	Supplier selection, sourcing cost improvement, sourcing risk management (Olson, 2015; Rozados and Tjahjono, 2014; Sanders, 2014, p.132)
	Manufacturing	Product Research and Development (R&D), production planning and control, quality management, maintenance and diagnosis (Meziane and Proudlove, 2000)
	Logistics/ Transportation	Intelligent transportation system, logistics planning and in-transit inventory management (Wegner and Küchelhaus, 2013)
	Warehousing	Storage assignment, order picking, inventory control (Rozados and Tjahjono, 2014)
	Demand management	Demand forecasting, demand sensing, demand shaping (Chase, 2016)
Level of data analytics		Descriptive, Predictive, Prescriptive (Saumyadipta et al., 2016, p.15)
BDA models		Visualisation, association, clustering, classification, regression, forecasting, semantic analysis, optimisation, simulation (Erl et al., 2016, p181)
BDA techniques		Association rule mining, clustering algorithms, support vector machine, linear/logistics regression, neural network, fuzzy logic, Naïve Bayes, text mining, sentiment analysis, feature selection, OLAP, statistics, to name a few.

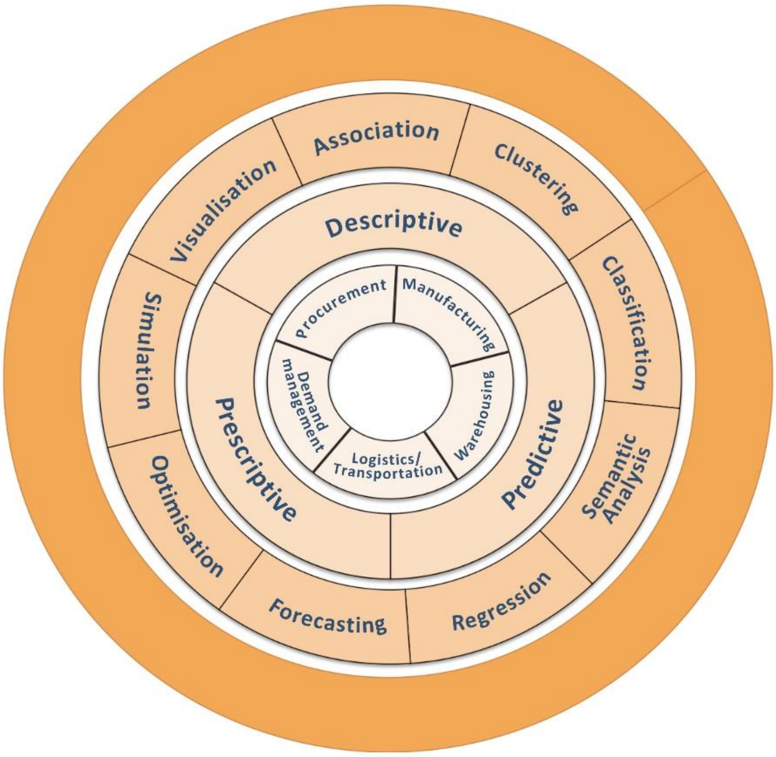


Figure 2.4 - Classification framework

In order to ensure the all-inclusive categorization of each articles being reviewed, there are some supplement categories, for example, 'General SCM', 'Mixed' and 'N/A. To avoid confusion, those categories are not presented in the graphical classification framework in Figure 2.4.

The first layer lies on the key functions of SCM. In the second layer, the author classified the BDA-SCM literature based on three levels of data analytics, namely descriptive, predictive and prescriptive. This taxonomy has been widely adopted in BDA studies as it reflects complexity of both BDA-applied problems and data analytics techniques (Delen and Demirkan, 2013; Duan and Xiong, 2015). Descriptive analytics are the simplest form of BDA, which describe what happened in the past; while predictive analytics are to predict future events, and prescriptive analytics refer to decision making mechanism and tools (Rehman et al., 2016). The third layer is nine most common types of BDA model in general (Erl et al., 2016, p181). The final is the layer of BDA techniques, which can be adopted from multiple data analytics disciplines such as data miming, machine learning, etc.

2.3 Material evaluation

2.3.1 Reviewing by SC functions

The distribution of the selected BDA studies in each SC function is presented in Figure 2.5, and the detailed classification based on key application topics is summarised in Table 2.2.

Overall, logistics/transportation and manufacturing have extremely dominated over the current literature on this topic, together taking up more than half of the publications. Research on the other three fundamental SC functions, namely warehousing, demand management and procurement are limited, but relatively well distributed.

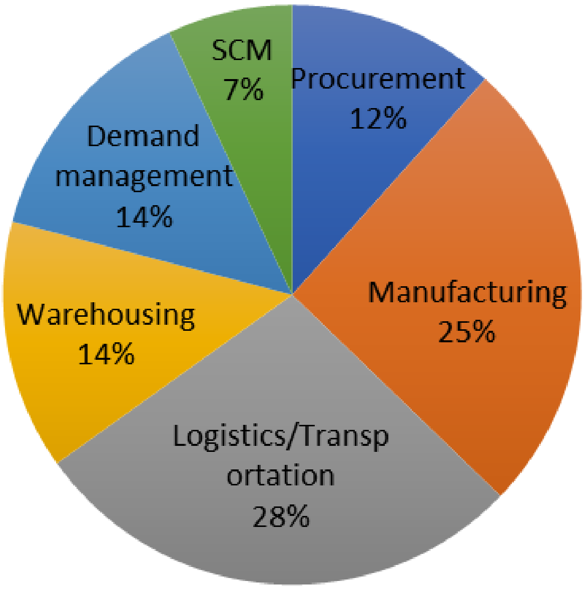


Figure 2.5 - Distribution of each SC function in BDA

Among five areas of SCM, logistics/transportation (24 out of 86 papers, 27.9%) is the most prevalent area where BDA is used to support decision-making. The majority of research papers in this area (15 papers, 62%) focus on using BDA to develop Intelligent Transportation System

(ITS), while BDA supporting logistics planning has increasingly gained attentions from recent research (7 papers or 29%). Only two papers (8.3%) are concerned about the use of BDA for inventory management during in-transit logistics process.

Another area in which BDA applications have been studied extensively is manufacturing. It takes up 25.6% of total publications. Out of the 22 manufacturing-related articles, more than half (12 papers, 54%,) are related to production planning and control, while production R&D and maintenance & diagnosis equally represent 27.3% of publications (6 papers each). Noteworthy, the use of BDA for quality control during manufacturing process is little discussed, appearing only in 3 papers.

In demand management literature, sensing current demand (7 papers) and shaping future demand (6 papers) are among the most prominent initiatives of BDA, while surprisingly, demand forecasting is seldom the focus of the study (only 3 papers).

In warehousing, BDA has been widely recognised to improve storage assignment (5 papers) and inventory management (5 papers), whilst the use of BDA to support order picking process is under-examined (3 papers).

BDA research focusing in procurement is well-balanced over three major issues: supplier selection (5 papers), sourcing improvement (4 papers) and sourcing risk management (4 papers).

While most studies in the literature examine the application of BDA to specific SC functions, the author found 6 papers (7% of total publications) that analyse BDA applications while considering SCs as multi-level interconnected networks. These papers address different SC issues concerning resilience, sustainability, risk management and agility.

Table 2.2 - Summary of literature by SC functions

SC Function	Key activity	Paper
Procurement	Supplier Selection	Choi et al., 2016; Huang and Handfield, 2015; Jain et al., 2014; Kuo et al., 2015; Mori et al., 2012
	Sourcing cost improvement	Ahiaga-Dagbui and Smith, 2014; Huang and Handfield, 2015; Kuo et al., 2015; Tan and Lee, 2015
	Sourcing risk management	Ghedini Ralha and Sarmento Silva, 2012; Huang and Handfield, 2015; Ling Ho and Wen Shih, 2014; Miroslav et al., 2014
Manufacturing	Product R&D	Bae and Kim, 2011; Do, 2014; Lei and Moon, 2015; Opresnik and Taisch, 2015; Tan et al., 2015; Zhang et al., 2017
	Production planning & control	Chien et al., 2014; Krumeich et al., 2016; Li et al., 2016; Wang and Zhang, 2016; Zhang et al., 2017; Zhong et al., 2015; Shu et al., 2016; Dai et al., 2012; Zhang et al., 2015; Zhong et al., 2015; Zhong et al., 2016; Zhong, Xu, et al., 2015
	Quality management	Krumeich et al., 2016; Wang et al., 2016; Zhang et al., 2017; Zhang et al., 2015
	Maintenance & diagnosis	Shu et al., 2016; Guo et al., 2016; Kumar et al., 2016; Zhang et al., 2017; Wang et al., 2016; Wang et al., 2015
Warehousing	Storage assignment	Chuang et al., 2014; Li, Moghaddam, et al., 2016; Tsai and Huang, 2015; Chiang et al., 2011; Chiang et al., 2014
	Order picking	Ballestín et al., 2013; Chuang et al., 2014; Li et al., 2016
	Inventory control	Alyahya et al., 2016; Hofmann, 2015; Hsu et al., 2015; Huang and Van Mieghem, 2014; Lee et al., 2016; Stefanovic, 2015
Logistics and Transportation	Intelligent Transportation system (ITS)	Cui et al., 2016; Li et al., 2015; Shi and Abdel-Aty, 2015; St-Aubin et al., 2015; Toole et al., 2015; Wang et al., 2016; Xia et al., 2016; Yu and Abdel-Aty, 2014; Zangenehpour et al., 2015; Dobre and Xhafa, 2014; Ehmke et al., 2016; Sivamani et al., 2014; Toole et al., 2015; Zhang et al., 2016; Hsu et al., 2015
	Logistics planning	Lee, 2016; Prasad et al., 2016; Yan-Qiu and Hao, 2016; Zhao et al., 2016; Shan and Zhu, 2015; Tu et al., 2015; Li et al., 2014
	In-transit inventory management	Ting et al., 2014; Delen et al., 2011
Demand management	Demand forecasting	Berengueres and Efimov, 2014; Chong et al., 2016; Jun et al., 2014; Li, Ch’ng, et al., 2016; Ma et al., 2014
	Demand sensing	Berengueres and Efimov, 2014; Chong et al., 2016; Fang and Zhan, 2015; He et al., 2015; Li, Ch’ng, et al., 2016; Salehan and Kim, 2016; Wang et al., 2014
	Demand shaping	Chong et al., 2016; He et al., 2015; Marine-Roig and Anton Clavé, 2015; Salehan and Kim, 2016; Schmidt et al., 2014
General SCM	Ong et al., 2015; Papadopoulos et al., 2017; Sheffi, 2015; Ting et al., 2014; Wu et al., 2017; Zhao et al., 2016, 2015	

2.3.2Reviewing by level of data analytics

The rationale of using this taxonomy is to examine the extent to which BDA is being used to support decision making processes, as well as understanding what types of SC problems being solved.

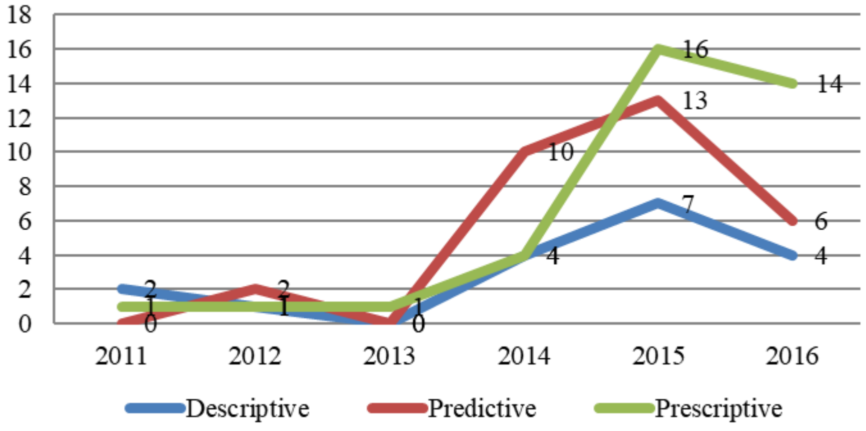


Figure 2.6 - Distribution of analytics level by year

Figure 2.6 depicts the popularity of each analytics type by year. Although the trend from 2011 to 2013 is underrepresented due to the insufficient number of BDA-SCM studies, it still can be seen that the majority of studies in this early stage used BDA for descriptive analytics, while predictive and prescriptive analytics had been little discussed. However, these minorities have changed drastically along with the upsurge in the publication level since 2013. Particularly, predictive analytics dominated in 2014, accounted for 55% of publications (10 out of 18 papers). Meanwhile, prescriptive analytics has been the fastest growing since 2014 and has become the most common type in 2015 with 16 out of 36 papers (44.4%). Descriptive analytics have also been aligning with these upward trends, but not remarkably soared like prescriptive analytics.

To reveal more insights from this taxonomy of BDA, the author further investigates how each level of analytics has been studied in each specific SC domain. Table 2.3 presents the result. Overall, prescriptive analytics is the most discussed type in the examined literature, taking up 43% of publications (37 out of 86 papers), while predictive analytics is just behind with 31 papers (36.1%), and finally, descriptive analytics (18 papers, 20.9%). In particular, the result found that manufacturing (11 papers) and logistics/transportation (14 papers) are those areas which have mainly contributed to the prominence of prescriptive analytics, thanks to the increasing adoption of various state-of-the-art systems such as Cyber Physical System (CPS), and Intelligent Transportation System (ITS). Meanwhile, predictive analytics is the most often used type in demand management (6 out of 12 papers) and procurement area (4 out of 10 papers). Accurate demand forecasting and early detection of various sourcing risks are among foremost applications of BDA-enabled predictive models in these two areas. It should be noted that prescriptive applications of BDA in demand management and procurement are seldom studied.

Table 2.3 - Level of analytics in each SC function

	Descriptive	**Predictive**	**Prescriptive**
Procurement	3	5	2
Manufacturing	4	7	11
Warehousing	3	3	6
Logistics/Transportation	4	6	14
Demand management	2	10	0
Other SC topics	2	0	4
Total papers	18	31	37

2.3.3 Reviewing by types of BDA models

The result of this review is summarised in Table 2.4. As discussed in Section 2.3.2, since the majority of reviewed papers focus on the prescriptive applications of BDA, the adoption of optimisation and simulation modelling to support decision-making seems to be natural. Optimisation is adopted in 18 papers in Table 2.4, but surprisingly enough, simulation is only applied in 6 papers. Route optimisation and logistics planning in logistics/transportation areas (8 papers) have mainly attributed to the dominance of optimisation, whereas in contrast, there is no paper found in examined literature that uses simulation in this area.

Table 2.4 - Distribution of articles by BDA models

	Procurement	**Manufacturing**	**Warehousing**	**Logistics/ Transportation**	**Demand management**	**General SCM**	**Total papers**
Optimisation	1	4	4	10			**19**
Classification	2	5		3			**11**
Mixed/ Others		4		3		3	**10**
Association	2	1	3	2	2		**10**
Semantic analysis	1	1			6	1	**9**
Forecasting	1		2	2	3		**8**
Simulation	1	3	2	1			**7**
Clustering	1			2	1	1	**5**
Regression		1	1	1			**3**
Visualisation		2					**2**

For the predictive level of analytics, classification is the most used BDA model in SCM context (11 papers). The model aims to classify a huge set of data objects into predefined categories, thereby generating prediction with high accuracy (Mastrogiannis et al., 2009). Classification has been largely employed in manufacturing (5 out of 11 papers, or 45%), logistics/transportation (3 papers, 27%), and procurement (2 papers, 11%). Other popular models for predictive analytics are semantic analysis (9 papers) and forecasting (8 papers). While the application scope of forecasting models is quite diversified, ranging from demand management, warehousing, and logistics/transportation to procurement, more than half of studies using semantic analysis (6 out of 9 papers or 66,7%) are still limited to demand management area.

At the descriptive level of BDA, association is the most used approach (11.6%, 10 out of 86 papers), which refers to the discovery of recurring and strong relationships among items in large-scale datasets. It should be noted that association is also the most diversified model in the BDA-SCM field as it can support decision making in every stage of SC process, from procurement, manufacturing, warehousing, logistics/transportation to demand management. Visualisation, surprisingly, is the least used model in descriptive analytics as well as in the

BDA-SC literature as a whole. There are only two papers in the examined literature studying this type of model as the main research focus.

Finally, a number of papers are classified into mixed/others models (10 papers, 11.6% of publications). Those papers fall into three SC functions, namely manufacturing (3 papers), logistics/transportation (3 papers), and general SCM (4 papers), as seen in the table.

2.3.4Reviewing by BDA techniques

This section focuses on what types of BDA techniques and algorithm have been used to develop BDA models discussed in Section 2.3.3. The result is summarised in Table 2.5.

Table 2.5 - Distribution of articles by BDA techniques

	Association	**Clustering**	**Classification**	**Regression**	**Forecasting**	**Semantic analysis**	**Optimisation**	**Simulation**	**Visualisation**	**Mixed**	**Total papers**
Association rule mining	8		1				4				**12**
Mixed*		1	2							7	**9**
Heuristic approach							7	1			**8**
Spatial-temporal visualisation							5	1	1		**7**
Decision tree	1		3		3						**7**
Support vector machine			6								**6**
Sentiment analysis						3			2		**5**
Logistic regression			1	3	1						**5**
Generic Algorithm							4				**4**
Neural network					3	1					**4**
Text mining		1	1			2					**4**
K-mean clustering		1	1		1			1			**4**
Time-series forecasting					3		1				**4**

Tables 2.5 highlights the prevalent use of some techniques to particular types of BDA models.

For instance, support vector machine (SVM) (6 papers) prevails in classification models, while heuristic approaches (7 papers) along with spatial/temporal-based visual analysis (5 papers) are key methods to develop BDA-driven optimisation models. In addition, there are some techniques that can be used flexibly to different types of BDA modelling. For example, K-means clustering algorithm is among the most adaptable techniques as it can be applied in clustering, classification, forecasting, and simulation model. Another versatile technique, and also the most frequently used technique in the BDA-applied SCM literature is Association rule mining (ARM). This method has been extensively studied in descriptive association model (7 papers) but recently, is increasingly used to facilitate more complicated analytics models in predictive and prescriptive level such as classification (1 paper) and optimisation (4 papers).

2.4 In-depth content analysis based on each SC function

2.4.1 Procurement

BDA in procurement lies on supply management, spend analysis and risk prediction, as presented in Table 2.6. Research on supplier selection and relationship development inherits conventional process where a number of selection criteria are set up in first place. The role of BDA is to enable these criteria to be able to expand to a considerable multifaceted scale, i.e. multi-dimensions and multi-layers of multi-criteria, which reflects the complexity of supply management (Choi et al., 2016; Jain et al., 2014; Mori et al., 2012). The supplier selection often accompanies with effort of reducing cost through allocating order quantity between key suppliers (Kuo et al., 2015), improving purchasing process (Tan and Lee, 2015), predicting uncertainties (Ahiaga-Dagbui and Smith, 2014) and estimating pricing risk (Huang and Hanfield, 2015). Moreover, BDA also expedites risk management through established warning system of abnormal events (Ling Ho and Wen Shih, 2014), detecting suppliers cartel (Ghedini Ralha and Sarmento Silva, 2012) and corruption (Miroslav et al., 2014).

It is noted that developing a knowledge base (KB) for decision-making and risk identification is a common practice. The KB serves as a reference and benchmarking to various supplier selection criteria development and risk detection alteration. Nevertheless, research on developing a comprehensive KB itself is rather scarce.

In terms of techniques, BDA in supplier selection has been using ‘static’ historic data to help make decision. There seems to lack of research of using real-time data that can enable dynamic contracting and relationship between customer and supplier in line with the changing business environment.

Table 2.6 - Summary of literature on BDA in procurement

Article	Research Type	Supplier Selection	Sourcing cost improvement	Sourcing risk management	Level of analytics	BDA model type	BDA techniques
Choi et al. (2016)	Model	*			Prescriptive	Simulation	Fuzzy cognitive mapping
Mori et al. (2012)	Model	*			Predictive	Classification	Support vector machine
Jain et al. (2012)	Model	*			Descriptive	Association	Fuzzy association rule
Huang and Hanfield (2015)	Model	*	*	*	Descriptive	N/A	SC maturity rating model
Kuo et al. (2015)	Model	*	*		Prescriptive	Optimisation	Association rule, artificial immune network, Particle swarm optimization
Melvin Tan and Wee-leong (2015)	Model		*		Descriptive	Clustering	Text mining, K-mean clustering
Ahiaga-dagbui et al. (2014)	Model		*		Predictive	Forecasting	Artificial neural network
Ho and Shih (2014)	Platform			*	Predictive	Classification	Association rule mining, decision tree
Ralha and Silva (2012)	Model & Platform			*	Predictive	Association	EM Clustering, association rule
Miroslav et al. (2014)	Model & Platform			*	Predictive	Semantic analysis	Meta-Model

2.4.2Manufacturing

BDA in manufacturing is to streamline manufacturing process by identifying core determinants that influencing SC performance as a whole, and then taking actions to continuously improve them. This has been studied in four aspects: product research and development (R&D), production planning and control (PPC), quality management, diagnosis and maintenance, as presented in Table 2.7.

Product R&D heavily relies on extracting various forms of customer feedback (Kwon and Kim, 2011; Lei and Moon, 2015). Through BDA, the result of these feedback analyses can lead to enhanced innovation capability (Tan et al., 2015) and monitoring on-going product development performance (Do, 2014). It also exploits the new revenue generation streams and opens opportunities for cost reduction (Opresnik and Taisch, 2015). Yet, the production quality can be seen as one of key performance indicators (KPIs). However, BDA application on this area is only partially explored in literature, and mostly discussed along with other activities in manufacturing process such as production planning and control, resources allocation and process monitoring (Zhang et al., 2015; Krumeich et al., 2015). Studies that focus particularly on BDA-driven quality control are very few (Zhang et al., 2015).

Applying BDA in diagnosing machine fault and planning for maintenance is another area that has been gained rising attentions of academic. It is expected that the system would be able to automatically detect the failure and be capable of taking action without human intervention (Zhang et al., 2015; Kumar et al., 2016; Wang et al., 2015; Shu et al., 2015). The fault refers not only machines but also workers' abnormal behaviours (Guo et al., 2016b). It is interesting to note that this is one of few areas for which the BDA techniques are developed to accommodate distributed agents such as cloud-based platform, machines, conveyers, products, to name a few.

Finally, BDA in production planning and control dominates over the above three aspects in manufacturing management. In order to meet customer demand, production planning and control must be aligned, which requires accurate forecasting in terms of production order arrivals (Zhong et al., 2015) and production cycle time (Wang and Zhang, 2016). By doing this, RFID-enabled production scheduling can allocate resources effectively (Wang and Zhang, 2016; Zhong et al., 2015) and release the shop floor planning (Zhong et al., 2015; Wang and Zhang, 2016; Lan, et al., 2015). In many cases, the planning and control executes through RIFD/sensors-enabled real-time intelligent cloud manufacturing system (Zhong et al., 2015; Zhong et al., 2015; Zhong et al., 2015; Helo and Hao, 2017). The applied BDA techniques are rather diversity, for instance, event-based processing prediction (Krumeich et al., 2015), heuristic based optimisation (Zhong et al., 2015), or bespoke algorithm developed for a specific platform (Zhong et al., 2015).

Table 2.7 - Summary of literature on BDA in manufacturing

Article	Research types	R&D	PPC	Quality management	Diagnosis & Maintenance	Level of analytics	BDA model type	BDA techniques
Zhang et al. (2016)	Platform	*	*	*	*	Prescriptive	Mixed/others	Mixed
Tan et al. (2015)	Model	*				Prescriptive	Optimisation	Deduction graph
Lei and Moon (2015)	Model & Platform	*				Prescriptive	Simulation	K-means clustering, AdaBoost classification
Do (2014)	Platform	*				Descriptive	Visualisation	Online analytical processing
Bae and Kim (2011)	Model	*				Descriptive	Association	Association rule, decision tree
Opresnik and Taisch (2015)	Theory	*				Descriptive	N/A	N/A
Zhong et al (2015)	Model		*			Prescriptive	Optimisation	Heuristic approach
Wang and Zhang (2016)	Model		*			Predictive	Classification	Mixed
Li et al. (2016)	Model		*			Prescriptive	Optimisation	Heuristic approach
Chien et al. (2014)	Model		*			Predictive	Classification	K-mean clustering, decision tree
Helo and Hao (2017)	Theory		*			Prescriptive	Optimisation	Mixed
Krumeich et al. (2015)	Theory		*	*		Prescriptive	Mixed/others	Mixed
Zhang et al. (2015)	Model			*		Prescriptive	Simulation	Spatial-temporal visualisation
Wang et al. (2016)	Model			*	*	Prescriptive	Simulation	Agent based simulation
Kumar et al. (2016)	Model				*	Predictive	Classification	Support vector machine
Wang et al. (2015)	Theory				*	Predictive	Classification	Support vector machine
Guo et al. (2016)	Platform				*	Predictive	Semantic analysis	Text mining
Shu et al. (2016)	Theory		*		*	Predictive	Classification	Entropy method, artificial immune network
Zhang et al., (2015)	Model & Platform		*			Prescriptive	Optimisation	Spatial-temporal visualisation, dynamical optimisation
Dai et al. (2012)	Platform		*			Prescriptive	Mixed/others	Mixed
Zhong et al. (2015)	Model		*			Prescriptive	Mixed/others	Mixed
Zhong et al. (2015)	Platform		*			Descriptive	Visualisation	Spatial-temporal visualisation
Zhong et al. (2015)	Model & Platform		*			Predictive	Regression	Curve fitting

2.4.3 Logistics and Transportation

Table 2.8 - Summary of literature on BDA in Logistics/Transportation

Article	Research type	Intelligent transportation system	Logistics planning	In-transit inventory management	Level of analytics	BDA model type	BDA techniques
St-Aubin et al. (2015)	Model	*			Descriptive	Clustering	K-means clusters
Shi and Abdel-aty (2015)	Model	*			Predictive	Regression	Random forest, Bayesian logistic regression
Yu and Abdel-aty (2014)	Model	*			Predictive	Classification	Support vector machine, logistic regression
Zangenehpour et al. (2015)	Model	*			Predictive	Classification	Support vector machine
Wang et al. (2015)	Platform	*			Prescriptive	Optimisation	Fuzzy logic, generic algorithm
Xia et al. (2015)	Model	*			Predictive	Forecasting	K-Nearest neighbour
Li et al. (2015)	Model	*			Predictive	Forecasting	Granger causality, Lasso regression
Cui et al. (2015)	Model	*			Descriptive	Association	Descriptive statistics
Walker and Strathie (2015)	Model	*			Descriptive	Clustering	Mixed
Dobre and Xhafa (2014)	Platform	*			Prescriptive	Mixed/others	Mixed
Toole et al. (2015)	Platform	*			Prescriptive	Optimisation	Heuristic approach
Fabian et al. (2016)	Model	*			Prescriptive	Optimisation	Heuristic approach
Zhang et al. (2016)	Model	*			Prescriptive	Optimisation	Spatial-temporal visualisation; heuristic approach
Hsu et al. (2015)	Platform	*			Prescriptive	Mixed/others	Mixed
Sivamani et al. (2014)	Platform	*			Prescriptive	Optimisation	OWL (Ontology Web Language)
Mehmood et al. (2017)	Model		*		Prescriptive	Optimisation	Markovian approach
Lee (2016)	Model		*		Prescriptive	Optimisation	Association rule, if-then prediction, generic algorithm
Yan-Qiu and Hao (2016)	Model		*		Prescriptive	Optimisation	Association rule, time-series forecasting, simulation
Zhao et al. (2016)	Model		*		Prescriptive	Optimisation	Mixed
Prasad et al. (2016)	Model		*		Descriptive	Association	Resource Dependence Theory

Shan and Zhu (2015)	Model		*		Prescriptive	Optimisation	Spatial-temporal visualisation, heuristic approach
Tu et al. (2015)	Model		*		Prescriptive	Optimisation	Spatial-temporal visualisation, generic algorithm
Li et al. (2014)	Model		*		Predictive	Classification	Support vector machine
Ting et al. (2014)	Model			*	Prescriptive	Simulation	Association rule, Dempster's Rule of combination
Delen et al. (2011)	Platform			*	Prescriptive	Mixed/others	Mixed

Overall, the study of BDA in logistics/transportation area is quite extensive but rather unbalanced. As can be seen in Table 2.8, transportation management prevails among this area with more than half of research (15 out of 24 papers, 62.5%). Meanwhile, BDA-driven logistics planning is still under-investigated (7 papers, 29.2 %) and in-transit inventory management is seldom discussed.

As regarding to transportation management, the ultimate objective of BDA adoption is to develop Intelligent Transportation System (ITS) that allows real-time traffic operation control and proactive safety management. Traffic operation efficiency can be improved either by accurately and timely prediction of short-term traffic flow demand (Xia et al., 2015; Li, Su, et al., 2015) or by using smart routing to avoid traffic congestion (Zhang et al., 2016; Fabian et al., 2016) whilst safety is mainly studied through crash analysis (Yu and Abdel-Aty, 2014; St-Aubin et al., 2015; Zangenehpour et al., 2015). It should be noted that Shi and Abdel-Aty (2015) is the only paper found in the examined literature that highlights the importance of jointly controlling and improving both traffic operations and safety simultaneously. In addition, most of studies above develop BDA-based models in static settings, whereas real-time decision-making support such as dynamic routing optimization or proactive traffic monitoring are only conceptually discussed in platfrom-based papers (Dobre and Xhafa, 2014; Wang et al., 2015; Toole et al., 2015; Hsu, Yang, et al., 2015, Sivamani et al., 2014).

Regarding to logistics planning, BDA can facilitate a range of strategic and operational decisions. For example, at strategic level, Tu et al. (2015) and Shan and Zhu (2015) analyzed large-scale GPS data to optimise facility location. Mehmood et al. (2017) used a Markov model to demonstrate how big data could be leveraged for optimise transport load sharing in smart cities to improve transport efficiency and reduce externalities. For operational capacity planning, Lee (2016) and Liu and Wang (2016) optimise order allocation and shipping assignment by extracting a sheer amount of customer location and consumption data for higher prediction accuracy of customer demand. Li et al. (2014) take advantage of massive historical data from detectors to accurately predict failures in rail operation, thereby optimizing maintenance scheduling. Noteworthy, the adoption of BDA approach in more holistic SC network design models that optimise both strategic and operational decisions simultaneously is still extremely limited (Zhao et al., 2016; Prasad et al., 2016).

To logistics and transportation planners, the major concern on inventory management is to define appropriate logistics plans in order to maintain the quality and safety of the product during in-transit process. From this viewpoint, BDA provides unprecedented opportunities for tracking, assessing and monitoring the product conditions in accordance with in-transit context, including temperature, vibration, moisture, light exposure, and humidity level. However, such decision support system (DSS) allowing in-transit inventory management is currently far less studied in literature, compared to other BDA-enabled logistics applications (Ting et al., 2014; Delen et al., 2011).

2.4.4 Warehousing

Efficiently handling and storing products and materials are the vital roles of warehousing, as summarized in Table 2.9. BDA applications in this supply chain function have focused particularly on material handling and layout zoning to maximizes space utilization, minimize distance travelled to fulfil orders and consequently minimize storage and material handling costs and risk of hazard events. A wide range of factors affect the way warehouses are managed, such as size and layout of storage and material handling systems, order picking policies, product features, order frequencies, demand trends, turnover rate, etc. (Chan and Chan, 2011). Data mining-based storage assignment methods have been used to process data on orders, products and customer constantly and automatically saved in warehouse management systems (WMS) and Enterprise Resource Planning (ERP) systems (Chuang et al., 2014; Chiang et al., 2011; Chiang et al., 2014; Li et al., 2016). In addition, mining of data collected through RFID enables the effective shelf space allocation and product positioning in retail shops considering customer purchase and browsing behaviours (Tsai and Huang, 2015).

While the use of BDA in storage assignment is quite extensively studies, its adoption in order pick is far less behind. Indeed, studies often discussed the advantages of BDA on order picking efficiency as a by-product of BDA-based optimal storage assignment (Chuang et al., 2014), while the study of how BDA can optimise order picking processes, such as order batching, routing, and sorting, is still scare (Ballestín et al., 2013; Alyahya et al., 2016).

A basic element of customer service is inventory availability. Inventory control has significant impacts across the whole SC process and directly dictates warehouse operations. Its primary task is to determine the right stock level that balances the inventory holding cost and the cost of lost sales. From this viewpoint, BDA has emerged as a key tool to support inventory managers. Indeed, the adoption of BDA stimulates collaborative planning, forecasting and replenishment (CPFR) implementation, which enables the real-time access, amalgamation and extraction of massive data from heterogeneous inventory points including procurement, production, distribution, and point of sale in order to produce fast, accurate, and reliable inventory replenishment predictions (Prajogo and Olhager, 2012). Previous research has explored the combination of traditional inventory planning and control methods with other data sources, such as clickstream data (Huang and Van Mieghem, 2014), manufacturing and customer behaviour data (Hsu et al. 2015) and counterpart stores and franchisers data (Stefanovic, 2015; Lee et al., 2015). The latter two works demonstrated the power of information sharing to improve inventory replenishment. By consolidating retail store data (Stefanovic 2015) and developing and exploring a cloud-based environment for real-time data sharing between franchisers (Lee et al. 2015), responsive replenishment systems were created.

Although it is well established that supply chain dynamics such as bullwhip effect, backlash effect and ripple effect are driven by different inventory control policies, the only study to cover this subject was Hofmann (2015). In a theoretical analysis, he explored an inventory control model to determine which characteristics of Big Data (volume, variety and velocity) are most likely to mitigate the bullwhip effect.

Table 2.9 - Summary of literature on BDA in warehousing

Article	Research type	Storage assignment	Order picking	Inventory control	Level of analytics	BDA model type	BDA techniques
Chuang et al. (2014)	Model	*	*		Descriptive	Association	association rule
Chiang et al. (2011)	Model	*			Descriptive	Association	Association rule
Chiang et al. (2014)	Model	*			Descriptive	Association	Association rule
Tsai and Huang (2015)	Model	*			Prescriptive	Optimisation	Association rule, combinatorial optimisation
Li, Moghaddam et al. (2016)	Model	*			Prescriptive	Optimisation	Association rule, generic algorithm
Ballestín et al. (2013)	Model		*		Prescriptive	Simulation	Heuristic approach
Alyahya et al. (2016)	Model & Platform		*		Prescriptive	Optimisation	Logistic regression
Huang and Van Mieghem (2014)	Model			*	Predictive	Regression	K-means clustering, decision tree, neural networks
Stefanovic (2015)	Model & Platform			*	Predictive	Forecasting	Fuzzy logic
Lee et al. (2015)	Model & Platform			*	Prescriptive	Optimisation	Neural networks
Hsu, Lin, et al. (2015)	Model			*	Predictive	Forecasting	Z-transforms and system dynamics simulation
Hofmann (2015)	Model			*	Prescriptive	Simulation	Spatial-temporal visualisation, heuristic approach

2.4.5 Demand management

BDA have been applied in demand management for demand forecasting, sensing and shaping (see Table 2.10). Demand forecasting normally requires predictive analytics using time-series approaches, auto-regressive methods and associative forecasts (Wang, Gunasekaran, et al., 2016). The contribution of BDA in demand forecasting has been on the association of time-series methods with other product- and market-related attributes. For instance, Ma et al. (2014) developed a Demand Trend Mining algorithm, which combines time series forecasting with product design attributes. Another example is the correlation found between time-series data on online search traffic information and oil price, consumer price index and product market share (Jun et al., 2014). Hence, search traffic can be used as an additional associative forecasting resource for certain products. In Berengueres and Efimov (2014) work, regression methods were also used to integrate data in loyalty card to improve demand predictions.

Demand sensing in combination with BDA techniques has enabled companies to incorporate detailed short-term and real-time demand data into their forecasts. For instance, some studies have demonstrated how online reviews (Chong et al., 2016; Fang and Zhan, 2015; Li et al., 2016) and social media data (He et al., 2015) can be used to determine the predictors of product sales in both e-commerce and retail business. By using sentiment analysis and text clustering, these studies were able to translate unstructured comments to describe and predict customer behaviour. Another research stream has made efforts to analyse the readership and helpfulness of online comments for both vendors and consumers (Salehan and Kim, 2015). Search traffic information, loyalty cards and mobile network data have also proven to be useful in sensing demand through customer segmentation (Berengueres and Efimov, 2014, Wang et at., 2014; Jun et al., 2014). Detecting customer behaviour and predicting market volatility have underscored the opportunity to sense and react in near real-time to changes in the demand.

After forecasting and sensing demand, BDA can be applied to shape demand, such as by price management (Schmidt et al., 2015), marketing and advertisement (Chong et al., 2016), managing online reviews (Salehan and Kim, 2015), long-term marketing campaigns and planning policies, improving branding (Marine-Roig and Clave, 2015), improving customer experience (He et al., 2015) and product life cycle management (Ma et al., 2014). Although demand shaping is one of the operational supply chain management strategies for effective capacity planning, previous research has taken more the marketing intelligence perspective.

Table 2.10 - Summary of literature on BDA in demand management

Article	Research type	Demand forecasting	Demand sensing	Demand shaping	Level of analytics	BDA model type	BDA techniques
Berengueres and Efimov (2014)	Model	*	*		Predictive	Forecasting	Decision tree, logistic regression
Jun et al. (2014)	Model	*			Predictive	Forecasting	Time-series forecasting
Ma et al. (2014)	Model	*			Predictive	Semantic analysis	Sentiment analysis, neural network
Li, Ch'ng et al. (2016)	Model		*		Descriptive	Association	Hierarchical multiple regression analysis
Wang, Tu et al. (2014)	Model		*		Predictive	Forecasting	Decision tree, automatic time-series forecasting

He et al. (2015)	Model		*	*	Descriptive	Clustering	Fuzzy c-means clustering
Marine-Roig and Clavé (2015)	Model			*	Predictive	Semantic analysis	Text mining, sentiment analysis
Salehan and Kim (2015)	Model		*	*	Predictive	Semantic analysis	Content analysis
Fang and Zhan (2015)	Model		*		Predictive	Semantic analysis	Sentiment mining
Chong et al. (2016)	Model & platform		*	*	Predictive	Semantic analysis	Sentiment mining
Schmidt et al. (2015)	Theory			*	Predictive	Semantic analysis	Sentiment analysis

2.4.6 General SCM

Table 2.11 - Summary of literature on BDA in "General SCM" papers

Articles	Research type	Level of analytics	BDA model type	BDA techniques
Sheffi (2015)	Theory	Prescriptive	Mixed/others	N/A
Wu et al. (2016)	Model	Descriptive	Clustering	Entropy method, quantitative transformation function
Papadopoulos et al. (2016)	Theory	Descriptive	Semantic analysis	Content analysis, factor analysis
Ong et al. (2014)	Platform	Prescriptive	Mixed/others	Text mining, dashboard visualisation
Zou et al. (2016)	Model	Prescriptive	Mixed/others	Kalman filter algorithm, median filter algorithm
Giannakis and Louis (2016)	Platform	Prescriptive	Mixed/others	N/A

The author found 6 papers that, instead of focusing on a single SC function, examined the application of BDA considering SCs as multi-level interconnected networks (Table 2.11). These papers have addressed different SC issues concerning resilience (Sheffi 2005; Papadopoulos et al., 2016), sustainability (Papadopoulos et al., 2016; Wu et al., 2016), risk management (Ong et al., 2014) and agility (Giannakis and Louis, 2016). However, most of them propose theoretical framework or use real-world cases to illustrate how companies are currently using sensors, social media, and event monitoring web-services to improve risk detection, response and traceability. Two notable exceptions are Wu et al. (2016) and Giannakis and Louis (2016). The former used BDA to transform social media data into manageable information so that companies can quickly respond to customer needs. Likewise, Giannakis and Louis (2016) created a multi-agent based SCM system that incorporates autonomous corrective control actions to answer to different SC partners' requirements. Finally, by using a completely different angle from all the mentioned papers above, Zou et al. (2016) provides technical solutions for improving real-time data processing and accuracy in interconnected SC network nodes.

2.5 Discussion

(1) What areas in SCM that BDA is being applied?

In logistics, transportation management prevails with particular focus on three fundamental functions of ITS, i.e. routing optimization, real-time traffic operation monitoring and proactive safety management. It is noteworthy that the BDA-driven routing problem is mainly studied in static environment based on historical databases (Ehmke et al., 2016; Zhang et al., 2016), while the use of BDA for dynamic routing optimisation in real-time context is only conceptualised in some theoretical platform-based paper such as (Sivamani et al., 2014; Hsu et al., 2015). Moreover, the application of BDA on logistics network planning has gained rising attentions recently, but is still under-examined in both strategic and operational levels (Zhao et al., 2017). Finally, the monitoring and control of product condition through sensors during in-transit process is seldom addressed (Delen et al., 2011; Ting et al., 2014).

Production planning and control is currently receiving the most research interest, and the application of BDA theories and tools on this topic is in a relatively mature stage (Wang and Zhang, 2016; Zhong et al., 2015). Although BDA adoption in product R&D and equipment diagnosis and maintenance is less often studied, papers in this area make a significant contribution to predictive and prescriptive analytics in manufacturing research (Lei and Moon, 2015; Wang, Zhang, et al., 2015; Wang and Zhang, 2016a; Zhang, Ren, et al., 2015). Noteworthy, research on BDA-enabled quality control during manufacturing processes is rather limited (Krumeich et al., 2016; Zhang et al., 2015).

With regards to warehousing operation, storage assignment and inventory control taking advantage of BDA are well-studied. However, inventory control-related dynamics, such as the Bullwhip effect, have just recently been theoretically discussed (Hofmann, 2015). Furthermore, few studies addressed order-picking problems in BDA-enabled warehousing (Ballestín et al., 2013; Chuang et al., 2014). The study of how BDA can optimise order picking processes, such as order batching, routing, and sorting, is still scare.

Studies of BDA in procurement area is evenly spread over the three major applications, i.e. supplier selection, sourcing cost improvement and sourcing risk analysis. BDA has been well-adopted to facilitate supplier selection process and recent research efforts have been made to integrate this activity with order allocation problems and to reduce sourcing cost (Kuo et al., 2015). In terms of sourcing risk management, most studies have only exploited the benefit of BDA to accurately detect procurement risk based on the massive supplier database, while models and DSS that provides proactive preventing actions are still lacking (Ghedini Ralha and Sarmento Silva, 2012; Miroslav et al., 2014).

The examined literature provides numerous contributions in terms of capturing demand changes in real-time. BDA can help in sensing demand behaviours to increase the agility and accuracy of demand forecasting (Fang and Zhan, 2015; Salehan and Kim, 2016; Wang et al., 2014). Another common application of BDA in demand management is shaping demand to be aligned with production and logistics capacity. However, current studies on this issue have taken a marketing intelligence perspective rather than an operational supply chain management perspective (Marine-Roig and Anton Clavé, 2015; Schmidt et al., 2015).

Finally, the review shows that recent research has increasingly recognised the importance of studying BDA with a holistic perspective cognisant of SC as a multi-level inextricably interlinked system. Most of those studies examined the SC integration in the context of SC resilience (Papadopoulos et al., 2015; Sheffi, 2015), sustainability (Papadopoulos et al., 2015; Wu et al., 2017), risk management (Ong et al., 2015) and agility (Giannakis and Louis, 2016). However, the research on this issue still strongly emphasises on theoretical development with the limited studies of advanced data mining modelling.

(2) What level of analytics is BDA used in SCM?

The rationale of this research question is to investigate the level of data analytics required in the SC application, as well as indicating the types of problem being solved.

In the trend analysis, the results show that prescriptive analytics is the most common and fastest growing in the BDA-driven SCM, which is closely followed by predictive analytics, while descriptive analytics are receiving less consideration. To be more specific, logistics/transportation, manufacturing, and warehousing domains are the major contributors of prescriptive analytics, thanks to the increasing adoption of various state-of-the-art systems such as Cyber Physical System (CPS) in Industry 4.0 (Wang et al., 2016), and ITS (Wang, Zhang, et al., 2015). On the other hand, predictive analytics are still the primary actors in demand management and procurement, especially for demand forecasting and sourcing risks detection while prescriptive analytics are still rarely discussed (Ghedini Ralha and Sarmento Silva, 2012).

(3) What types of BDA models are being employed in SCM?

Optimisation is the most popular approach when it comes to prescriptive analytics, especially in logistics and transportation area. As aforementioned, literature in logistics/transportation has little insights so far on real-time routing optimisation based on streamline data. On the other hand, the study of real-time optimisation appears to be quite mature in the manufacturing domain with the use of modelling & simulation to develop real-time production control system based on streamline context-aware data generated from tracking devices such as RFID (Babiceanu and Seker, 2016; Kumar et al., 2016b). It is highly possible for transportation controllers and warehouse operators to adapt the similar approach of modelling & simulation to optimise routing problem in real-time, as suggested in (Wang et al., 2016).

Classification is the most common approach in predictive analytics level and has been widely applied in manufacturing to support production planning & control (Chien et al., 2014; Wang and Zhang, 2016a) and equipment maintenance & diagnosis (Kumar et al., 2016; Shu et al., 2016; Wang et al., 2016). This type of BDA model also plays a key role in logistics/transportation (Li, Parikh, et al., 2014; Yu and Abdel-Aty, 2014; Zangenehpour et al., 2015) and procurement research (Ling Ho and Wen Shih, 2014; Mori et al., 2012) but apparently, current studies in those areas have not been fully exploited the advantages of classification.

Another popular model for predictive analytics is semantic analysis, but its scope of application is still considerably limited to demand sensing. It could be beneficial for future research to extend the application of this approach to more SC and operational management such as fraud detection (Miroslav et al., 2014) and behaviour-based safety analysis (Guo et al., 2016a).

For descriptive analytics, association is the most widespread as it has been applied for every stage of SC process, from procurement (Ghedini Ralha and Sarmento Silva, 2012; Jain et al., 2014), manufacturing (Bae and Kim, 2011), warehousing (Chiang et al., 2011, 2014; Chuang et al., 2014), logistics/transportation (Cui et al., 2016), to demand management (Jin et al., 2016). Noteworthy, visualisation model is rarely considered as the main focus of a study (Zhong et al., 2016; Zhong et al., 2015), but is commonly used as a complement to other advanced data mining models (Shan and Zhu, 2015; Zhang et al., 2016).

Finally, the review classified 10 papers (11.6% of total papers) under “mixed/others” models. Most of those papers focus on the intelligent DSS that enables real-time control of the entire operational process in manufacturing (Dai et al., 2012; Krumeich et al., 2015; Zhang, Ren, et al., 2015; Zhong et al., 2016), logistics/transportation (Delen et al., 2011; Dobre and Xhafa, 2014; Hsu, Lin, et al., 2015) and SC agility (Ong et al., 2015; Papadopoulos et al., 2015). Not surprisingly, mathematical models are missing in many of those papers since such state-of-the-art systems normally require a complex mixture of various models and techniques from a number of data mining disciplines.

(4) What types of BDA techniques are being used in SCM context?

There is a wide range of BDA techniques and algorithms that have been used in SCM context. Some of them are prevalent to particular modelling approaches, for example, SVM in classification models, heuristics approach in optimisation models and neural network in forecasting models. The review also identifies some versatile techniques that can be adapted to different types of models. For instance, k-means clustering algorithm is among the most adaptable techniques as can be adopted in clustering (St-Aubin et al., 2015; Tan and Lee, 2015), classification (Chien et al., 2014), forecasting (Stefanovic, 2015), and modelling & simulation (Lei and Moon, 2015). In those studies, K-means is often performed in the initial phase of data analytics process to partition the raw heterogeneous datasets into more homogenous segments.

Studies find out that advanced data mining techniques such as decision trees and neural network would develop more accurate predictive models by leveraging the result of cluster analysis (Krumeich et al., 2015; Lei and Moon, 2015; Stefanovic, 2015).

However, other than for descriptive analytics, scholars have increasingly used this method to facilitate more complicated analytics in predictive and prescriptive level. As yet, the author only found one paper, (Ling Ho and Wen Shih, 2014), conceptualise the notion of using ARM along with decision tree to develop the highly accurate prediction models for procurement risk. However, the mathematical model and algorithm to put this idea into practise is still missing. Interestingly, although the predictive application of ARM is little discussed in literature, its contribution to prescriptive analytics seems to be more recognised. Indeed, the author found 4 papers incorporating ARM with optimisation models to effectively solve allocation problems in various SC areas. For example, Li et al. (2016) use ARM and Generic algorithm (GA) to optimise storage assignments, thus enhancing order-picking processes. Tsai and Huang (2015) optimise shelf space allocation by using ARM, sequential pattern mining and combinatorial optimisation approach. For logistics and transportation planning, Lee (2016) use ARM to extract purchase patterns and perform if-then-else rules to predict customer purchase behaviour, thus proposing GA approach to optimise anticipatory shipping assignment. Finally, ARM can also be used in the hybrid optimisation problems of supplier selection and order allocation (Kuo et al., 2015).

Not surprisingly, there are a large number of papers under the “mixed” category (i.e the combination of more than three different methods) since there is no single technique that is fully capable of managing the complex and diverse nature of BD (Chen and Zhang, 2014).

2.6 Future directions

Those findings discussed above suggest some future directions to capitalise the research development of BDA applications in the SCM context.

(D1) Further investigation of BDA application to SC function level

The review suggests a number of research gaps in each SC function. For example, quality control in manufacturing, dynamic vehicle routing and in-transit inventory management in logistics/transportation, order picking and inventory control system in warehousing, demand shaping in SC and operational research, procurement are some of those areas that are currently much less discussed.

(D2) Functional alignment strategy for the horizontal integration of BDA-driven SC

SCM is the multi-level process of which functions are all interlinked. Hence, fragmental efforts of BDA adoption to only one or two functions will not yield any significant, long-lasting competitive advantage. To avoid such fragmented efforts, the entire SC should be horizontally integrated by aligning BDA applications in different functions effectively. For example, production and logistics planning could incorporate with real-time demand sensing for cost reduction and higher service level. Indeed, alignment dissolves the boundary across functions.

To facilitate the horizontal integration throughout the SC, future BDA research should focus more on cross-functional problems such as vehicle routing and facility location, supplier selection and order allocation, demand-driven storage assignment and order picking.

(D3) Three levels of analytics should be equally examined

As aforementioned, current research focuses more on prescriptive analytics than descriptive and predictive analytics. Nonetheless, the application of BDA in any subjects, not just SCM, is always a linear process. In this process, the performance of prescriptive analytics would heavily rely on those of descriptive and predictive analytics as they dictate the value of critical parameters in prescriptive models (Duan and Xiong, 2015). To catalyse the rapid progression of BDA application in SCM, future research should balance the focus to all three levels of analytics.

(D4) Combining different data analytic techniques to develop more advanced and adaptive BDA models for DSS

Literature review has identified a number of BDA models commonly used in SCM applications as well as popular and versatile BDA techniques to build those models. Dynamic optimisation and simulation modelling should be further investigated in the context of BDA as they are baseline approaches for prescriptive analytics and DSS.

Moreover, although literature has extensively adopted visualisation techniques as supplement techniques to predictive and prescriptive models, little attention has been paid on improving data visualisation techniques. Future research should call for this gap because visualising of complex BD would expedite decision making.

(D5) Application of BDA on closed-loop supply chain management

It is rather surprising that research on applying BDA on reverse logistics and closed-loop supply chain (CLSC) is scarce. This might be due to the fact that collecting data for used products is extremely hard, which hinders introducing BDA into CLSC management. Nevertheless, the development of new technologies such as Internet of Things, machine-to-machine, would be able to overcome this barrier. The BDA that has already been applied in product life cycle design and assessment (Ma et al., 2014; Song et al., 2016) would be useful for predicting product returns and estimating the return quality. This is important for capacity planning and remanufacturing scheduling in a reverse logistics system.

Managing a CLSC has always been challenging due to the uncertainties and possible conflicting goals, i.e. profit vs. environment vs. social wellbeing. In this sense, big data would be useful in understanding people perception, devising multi-KPIs, monitoring operational process, and then taking corresponding actions. To achieve this, developing knowledge database, tools and techniques of BDA must be on the research agenda.

(D6) BD-driven business models in SC

Big data revolutionises SC business models. On the one hand, it shorten the supply chain layers; On the other hand, it expands revenue streams from exiting products to servitization, and creates new revenue streams from entirely new (data) products (Opresnik and Taisch, 2015). Nevertheless, the ecosystem that supports new business model is underdeveloped, the enablers and obstacles of the new business models remain unclear.

This leads to the research questions that the author proposes here: (1) what are the big data strategies in SCM; (2) how to increase the VALUE of big data, the most important five Vs; (3) how various stakeholders contribute to adding value of big data and what is the revenue sharing mechanism among the stakeholders in SC; (4) what is the dynamic impact of new business model on SC performance as whole; (5) what are the tipping points that transfer a conventional business model to a big data-driven business model.

(D7) New tools and BDA techniques for distributed SC and distributed computation

Cloud computing, countless sensors around us, distributed service resources, and distributed operational processes generate voluminous amounts of data. Coordinating a *distributed* SC and managing the complex procedures of different BDA are challenging (Li et al., 2016).

The review suggests that majority of current research have been focusing on one-location one-computer scenario. This doesn’t reflect the reality of *distributed* systems. The author calls for research on developing SC system-wide feedback and coordination based on BDA to optimise system performance (Wang et al., 2016). It is anticipated that the framework and operational mechanism of nowadays smart factory would be scalable to entire SC. This SC system should be self-organised reconfiguration and big-data-based feedback and coordination without or with very limited human intervention. To achieve this, apart from hardware and infrastructure, future research should develop more efficient data-intensive techniques and technologies (Chen and Zhang, 2014).

2.7 Updated literature review after 2017 until 2019

As the literature review is conducted in the first year of the PhD, hence the findings above are based on the literature collected up to 2017. This section is the updated literature review between January 2018 and September 2019.

By carrying out the literature searching and selection process in Section 2.2.1 with the publication time period between 2018 and 2019, there are 24 papers that closely fit the scope of this literature, which suggest that BDA in SCM has still been actively studied. Indeed, since the literature review of this thesis was published in 2017, it has gained a fast growing attention among the scholar community with 57 citations to date.

The new 24 papers are summarised in Table 2.12 in accordance with the classification framework in the aforementioned literature review. Especially, the fact that all of these papers are in line with one or some of the research directions identified in section 2.6 above, which somewhat validates the result of the literature review.

Table 2.12 - Summary of updated literature between 2018-2019

Article	Research type	Research direction[1]	Research problem	SCM area	Level of analytics	BDA model type	BDA techniques
Ijadi Maghsoodi et al. (2018)	Model	D1, D4	Supplier selection	Procurement	Prescriptive	Optimisation	K-mean clustering, Multiple criteria decision analysis
Su and Chen (2018)	Model	D1	Sourcing risk management	Procurement	Predictive	Classification	Text mining, Ontology
Wang, Wang, et al. (2018)	Model	D1, D4	Logistics planning	Logistics & Transportation	Descriptive	Visualisation	Incremental-Learning Algorithm
Chang et al. (2018)	Model	D1, D4	Intelligent transportation system	Logistics & Transportation	Prescriptive	Optimisation	Fuzzy inference, Particle swarm optimization
Zhan and Tan (2018)	Model	D6, D7	Competence set analysis	General SCM	Prescriptive	Optimisation	Deduction graph-based optimisation
Singh et al. (2018)	Model & platform	D4, D7	Supplier selection	Procurement	Prescriptive	Optimisation	Cloud computing
Gohar et al. (2018)	Platform	D7	Intelligent transportation system	Logistics & Transportation	Prescriptive	Optimisation	Mixed
Kaur and Singh (2018)	Model	D2	Supplier selection, order allocation	Procurement & Logistics	Prescriptive	Optimisation	Heuristic-based optimisation
Zhao et al. (2018)	Model	D1, D4	Intelligent transportation system	Logistics & Transportation	Predictive	Forecasting	Mixed
Zhu (2018)	Theory & platform	D4, D7	Logistics planning	Logistics & Transportation	Prescriptive	Optimisation	Mixed
Jiao et al. (2018)	Model	D5	Closed-loop SC	Logistics & Transportation	Prescriptive	Optimisation	K-L divergence, Distributionally robust optimization
Srinivas and Ravindran (2018)	Model	D1, D6	Logistics planning	Logistics & Transportation	Prescriptive	Optimisation	ML-based classification, Multiple criteria decision analysis

Yu et al. (2019)	Model	D4	Demand forecasting	Demand management	Predictive	Forecasting	Mixed
Kumar et al. (2019)	Model	D4, D6	Demand forecasting	Demand management	Predictive	Forecasting	Fuzzy-based, back-propagation neural network
Belaud et al. (2019)	Theory	D6	General SCM	General SCM	Descriptive	Visualisation	Mixed
Liu (2019)	Model & Platform	D4, D7	Demand sensing	Demand management	Predictive	Semantic analysis	Fuzzy gray situation decision-making
Cavalcante et al. (2019)	Model	D4	Sourcing risk management	Procurement	Prescriptive	Simulation	Linear regression, K-nearest neighbour, Simulation
Çalı and Balaman (2019)	Model	D4	Demand shaping	Demand management	Prescriptive	Optimisation	Sentiment analysis, Multiple criteria decision analysis
Vieira et al. (2019)	Platform	D4, D7	General SCM	General SCM	Prescriptive	Simulation	Mixed
Liu et al. (2019)	Model	D1	Demand sensing	Demand management	Predictive	Semantic analysis	Sentiment analysis
Kakatkar and Spann (2019)	Platform	D1, D7	Demand sensing	Demand management	Descriptive	Visualisation	Mixed
Dev et al. (2019)	Model	D6	SC Performance measurement	General SCM	Prescriptive	Simulation & Optimisation	Mixed
Tseng et al, (2019)	Model	D4	SC Performance measurement	General SCM	Descriptive	N/A	Fuzzy synthetic method, Decision making trial and evaluation laboratory method
Alic et al. (2019)	Platform	D7	Intelligent transportation system	Logistics & Transportation	Prescriptive	Mixed	Mixed

[1] Research directions are explained in Section 2.6

2.8 Summary

Based on the content analysis methodology of Mayring (2008), this literature review examined 86 journal papers to provide a full picture of where and how BDA has been applied within the SCM context. In particularly, the author developed a classification framework based on four research questions: (1) what areas in SCM that BDA is being applied, (2) what level of analytics is BDA used in these application areas, (3) what types of BDA models are used, and finally (4) what BDA techniques are employed to develop these models. Addressing these questions, the discussion has highlighted a number of research gaps and future directions for BDA applications to catalyse the research development of the topic.

One of the limitations of this literature review is the categorisation in classification framework remaining interpretative, which could lead to the concern on subjective bias. This is also one of the well-established issue of the content analysis method despite a number of validation being done (Seuring, 2013).

Chapter 3 Research Methodology

This chapter describe how this research was conducted by explaining the philosophical orientations within the paradigm adopted for it. In particularly, the constitutions of research paradigm are explained in Section 3.1. Data-driven science, the paradigm used in this thesis, is justified in Section 3.2.

3.1 Fundamentals of research paradigm

It is particularly important to understand the research paradigm of a study because it reflects the philosophical beliefs and principles that shape how the researcher sees the world, and how he/she interprets and acts within that world (Lather, 1986). In other words, research paradigm is the conceptual lens through which the researcher investigates the methodological aspects of the research project to determine what methods will be used and how data will be analysed (Kivunja and Kuyini, 2017). A research paradigm typically consists of three fundamental elements: ontology, epistemology and methodology (Guba, 1990).

3.1.1 Ontology

Ontology is a branch of metaphysics that deals with the nature of being, existence and reality. It concerned about such questions as: "*What is the nature of reality?", "Is the reality out there of an objective nature, or the result of one or more individual minds?*". According to Paterson et al. (2016), ontological assumptions can be broadly categorised into two configurations: objective and subjective. An ***objective*** perspective looks at reality as made up of solid ***objects*** that can be measured and tested, and which exists independently of our comprehension of it. By contrast, a ***subjective*** perspective looks at reality as made up of perceptions and interactions of the living subjects; in other words, our perceptions are what shape reality. Choosing which ontological assumption to use is crucial because it orients the researcher's thinking about the research problem, its significance, the approach to address research questions, and how to make meaning of the collected data (Kivunja and Kuyini, 2017).

3.1.2 Epistemology

Epistemology of a paradigm refers to the way in which the researcher acquires valid knowledge. It studies the nature of knowledge and justification by asking such questions as: "*Is knowledge something which can be obtained, or is it something which has to be personally experienced?*", "*What does it mean to say that we know something?",* and especially, "*How do we know what we know?*". These questions are vital because they help the researcher position himself/herself in the research context, so that discovering what else is new, given what is known (Kivunja and Kuyini, 2017).

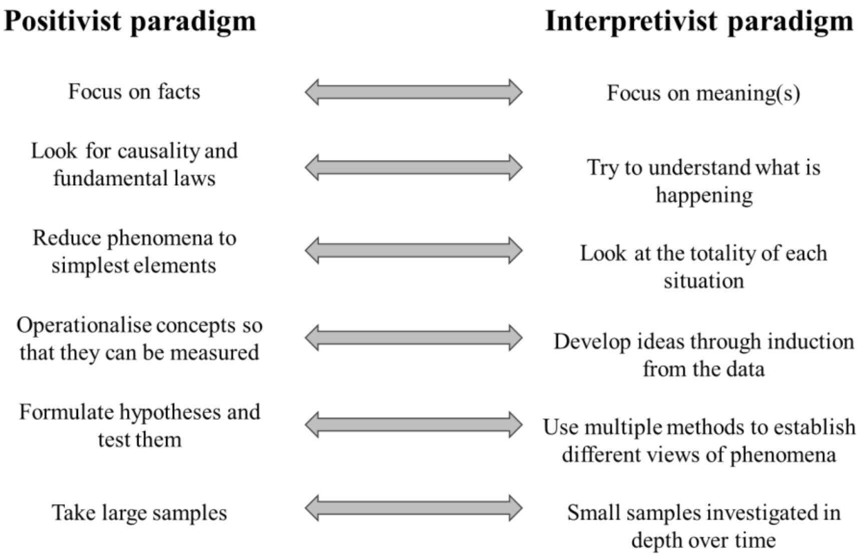
There are different types of epistemologies but positivism and interpretivism are the most widely used ones. ***Positivism*** is developed based on three assertions:

(1) Methodological procedures of natural sciences can be directly adapted to study human social behaviours;
(2) The research outcome of social sciences can take the forms of casual laws; and
(3) The results of social researches are value-free and unbiased (Paterson et al., 2016).

As opposed to the three assertions of positivism above, interpretivism paradigm holds that:

(1) There are essential differences between natural and social sciences, which stemmed from the different research aims – "explanation" versus "understanding". Interpretivists argued that social sciences seek to "understand" the social phenomena based on human experience, and therefore, the "causal-functional" approach used in natural sciences would not be applicable in social inquire;
(2) Due to self-consciousness of human being and freedom of choices, social scientists can only unveil "trends" rather than "laws"; and
(3) The results of social researches are value-laden and biased as they depend on the researcher's interpretation and beliefs (Paterson et al., 2016).

The more differences between positivism and interpretivism epistemology can be seen in Figure 3.1 below.



Positivist paradigm	Interpretivist paradigm
Focus on facts	Focus on meaning(s)
Look for causality and fundamental laws	Try to understand what is happening
Reduce phenomena to simplest elements	Look at the totality of each situation
Operationalise concepts so that they can be measured	Develop ideas through induction from the data
Formulate hypotheses and test them	Use multiple methods to establish different views of phenomena
Take large samples	Small samples investigated in depth over time

Figure 3.1 – Epistemologies with positivist and interpretivist influence

(Source: Paterson et al. (2016, p.68)

3.1.3Methodology

Methodology of a paradigm is an all-inclusive term which refers to the research strategies, methods, approaches, and procedures that are well-planned to investigate something (Keeves, 1997). For instance, data collection, participant identifications, instrumental design, and data analysis, are all parts of the broad field of methodology. In short, it focuses on the question of "how the knowledge of the world is gained?" (Moreno, 1947). It is vital to understand the researcher's methodology because it provides as the basis and rationale behind every method choice and collections of theories, concepts and ideas (Bryman and Bell, 2015). The brief descriptions of common methodological choices are presented as below.

- **Research strategy**

According to Saunders et al. (2019), there are three distinctive approaches to theory development – deductive, inductive and abductive approach. Using the ***deductive approach***, theoretical structure is first developed to formulate tentative propositions or hypotheses. Then, the researcher carries out an empirical observation to collect data and information to evaluate the propositions or hypotheses. Finally, the theory is confirmed if the conclusion passes the test. The outcome of this approach is theory verification or falsification. On the contrary, the ***inductive approach*** starts by collecting data and observing the empirical reality to analyse, compare and classify the facts before a hypothesis is formed. The outcome of this approach is theory generation and building. The ***abductive approach*** also starts with data collection to explore a phenomenon, identify themes and patterns to develop a new or modify an existing theory which is subsequently tested and revised via additional data collection, if necessary. In short, rather than moving from theory to data (as in deductive approach) or data to theory (as in inductive approach), the abductive approach moves back and forth, so that combining both deduction and induction (Suddaby, 2006).

Saunders et al. (2019) suggests that following a highly structured scientific approach that focuses on law-like generalisability, quantification and testable hypotheses, the deductive approach is most likely to be informed by subjective ontology and positivism epistemology. Conversely, due to its inevitable involvement to human being and its emphasis on the subjective interpretations, the inductive approach is most likely to be underpinned by interpretivism epistemology. Meanwhile, due to its flexibility, the abductive approach can be underpinned by various paradigms such as pragmatism or critical realism (Saunders et al., 2019).

- **Data collection method**

There are two main research methods for data collection, which are quantitative and qualitative research. ***Quantitative research***, which is also described by the term "empiricism" or "positivism", uses numerical data to quantify or measure a phenomenon and generate findings. It describes, measures and test cause and effect relationships, and often take a deductive approach to test theories (Bryman and Bell, 2015). The commonly used techniques for quantitative data collection are survey, questionnaires and experiments. By contrast, ***qualitative research*** deals with non-numerical data to understand certain aspects of a social problem from multiple perspectives. It studies the empirical world from the viewpoint of the subject rather than from the researcher, and generally adopts an inductive approach to build theories. Commonly used techniques to collect qualitative data are interviews, case studies, ethnography and focus groups.

3.2 Data-driven science – a new paradigm in Big Data era

This thesis aims to develop quantitative modelling that uses BDA to address SCM problems. Therefore, the author adopts an objective ontology and positivist epistemology when developing the methodology framework.

As discussed above, the quantitative-based scientific research conventionally relies on a deductive approach which is theory-driven hypothesis testing. However, big data, coupled with new data analytical techniques such as machine learnings and artificial intelligence, has been engendering a radical shift to a new research paradigm of ***data-driven science*** with new epistemologies and methodologies (Kitchin, 2014). The shift has been described as going from theory-driven to data-driven, from model-based to model-free sciences, and from parametric to no-parametric modelling (Russell and Norvig, 2016).

Although data-driven science seeks to share the tenet of the scientific methods, it is more open to use the hybrid approach combining aspects of deduction, induction and abduction to enhance the understanding of a phenomenon (Hey et al., 2009). More particularly, it differs from the traditional, knowledge-driven deductive approach in that it seeks to generate the hypotheses and insights "born from the data" rather than "born from the theory". In other words, it seeks to integrate a mode of induction into the research design, though from the positivist viewpoint, the induction-based explanation is not the intended end-point. Hence, instead of following the exact process of deductive or inductive approach, it forms a new mode of hypothesis formulation before a deductive approach is employed. As such, the new epistemological strategy adopted within data-driven science is to apply some knowledge discovery or pattern mining techniques to identify potential research questions or hypotheses that are worthy of further examination and testing. Such methods of hypothesis generation and data analysis are more likely to be based on abductive reasoning (Kitchin, 2014).

Adopting the data-driven scientific approach and applying BDA techniques, this thesis is able to undertake much richer data analysis, explore and extract insights from large-scale, interconnected datasets, identify and tackle research problems in new and exciting ways, as well as stimulating interdisciplinary research which conjoins domain expertise. By this way, the research can lead to more holistic and extensive models and theories of the entire complex system rather than the constitutes of them (Kelling et al., 2009).

There are a wide range of methods and tools that can be used in BDA-based research. The systematic classification of BDA models and techniques are reviewed in Chapter 2 above. As BDA is an interdisciplinary research, its methodological approach is very flexible and often tailor-made for the specific research problem. The methodology of each data-driven framework proposed in this thesis will be discussed in detail in each chapter below.

3.3 Summary

This chapter describes the fundamentals of the research philosophies and paradigms. It also explains how the explosion in Big Data generation and advances in BDA techniques have shifted the established social science into a new research paradigm, data-driven science. The ontology and epistemology of data-driven science also serves as the basis to the development of the methodology used in this thesis.

Chapter 4 Data-driven demand prediction of remanufactured products

The literature review in Chapter 2 has found that while BDA has been applied in almost every aspect of SCM, there are some aspects that have not yet been tapped into. CLSC is one of them. Hence, this chapter aims to shrink this gap by proposing a detailed approach of how BDA can be adopted to support the data-driven demand forecasting of remanufactured products – one of the core issue in the CLSC literature. Furthermore, the proposed approach is comprehensible to practitioners, meaning it can address the black-box issue which currently prevents the adoption of ML in practice. The research is expected to open up a number of opportunities for future research to explore new applications of BDA in CLSC.[2]

The chapter is structured as follows. Section 4.1 introduces research backgrounds and motivation, reviews related works, specifies research aims and objective, as well as positioning research contribution. Section 4.2 describes the methodology framework, detailing the rigorous research process adopted for this research. Section 4.3 describes the data collection and the variables related to the prediction model. Section 4.4 details the data preparation for the prediction model. Section 4.5 explains how the predictive models are developed and validated. Section 4.6 covers the evaluation and deployment of the model, while Section 4.7 discusses the results and the potential insights for management. Section 4.8 deals with the robustness checking of the model. Finally, section 4.9 is the conclusion.

[2] Chapter 4 has been published during the third year of the author’s PhD. See the reference:
Nguyen, T. V., Zhou, L., Chong, A., Li, B. and Pu, X. (2019). Predicting Customer Demand for Remanufactured Products: A Data-Mining Approach. *European Journal of Operational Research*.

4.1 Introduction

4.1.1 Research background and motivation

Remanufacturing is now a multi-billion dollar industry, with product sales increasing by 15% (approximately US$43 billion) per year. The driving forces of this growth are the enormous economic and environmental benefits gained from remanufacturing used products rather than producing new ones. These benefits include: reducing production costs by 50%; using 70% less raw materials, cutting manufacturing emissions by 80%; reducing energy consumption by a maximum of 60%; and offering lower prices to customers (Wang and Hazen, 2016). However, remanufacturing is not a panacea for the achievement of a sustainable thriving business, as its viability hinges on a high degree of uncertainty, in terms of both return and demand. The return uncertainty derives from a lack of information about the timing, quantity and quality of returned products (Zhou et al., 2016). The demand uncertainty of remanufactured products stems from unobservable remanufacturing processes, such as cleaning, disassembling, inspecting and testing, which make it difficult for customers to evaluate quality (Tereyağoğlu, 2016). Accurate predictions of demand for remanufactured products are therefore required for effective remanufacturing/CLSC operations.

Compared to studies in the CLSC subject related to returned product acquisition management and the operational issues of remanufacturing, less progress is being made in demand prediction and marketing strategy for remanufactured products (Atasu et al., 2008). Atasu et al. (2008) argue that there is an urgent need to apply advanced forecasting techniques to accurately quantify the uncertain parameters in the CLSC, such as demand, price and return. They point out that the more understanding of market acceptance of remanufactured products would facilitate the development of a more sophisticated analytical model of remanufacturing/CLSC operations with high industrial relevance.

This research is motivated by these two research streams – *remanufactured product demand forecasting* and *marketing strategy* – with the aim of contributing to CLSC research from both operational and marketing perspectives. To achieve this, the research utilises a data-mining and ML approach, which predicts demand for remanufactured products with high accuracy, resulting in a practicable marketing strategy for management. Related studies are reviewed below.

4.1.2 Review on market development of remanufactured products

Research into the market development of remanufactured products has attracted increasing interest over the last decade. The topic is broad and includes areas related to sales channels (Yan et al., 2015); warranty strategies (Alqahtani and Gupta, 2017); pricing analysis (Abbey, Blackburn, et al., 2015); and revenue management (Ovchinnikov, 2011). Of the various sub-topics, the differences in consumer behaviour involving remanufactured products and new products is one of the most emergent subjects. A literature review on the study of consumer behaviour in relation to remanufactured products is presented below and is summarised in Table 4.1.

Table 4.1 - Summary of studies on market development of remanufactured products

Article	Variable description		No. of independent variables	Interaction between variables	Analysis model		Data source	
	Dependent	Independent			Linear	Non-linear	Primary	Secondary
Wang, Wiegerinck, Krikke, & Zhang (2013)	Purchase Intention	Purchase Attitude, Perceived Risk, Perceived Benefit, Product Knowledge	4		✓		✓	
Jiménez-Parra, Rubio, & Vicente-Molina (2014)	Purchase Intention	Purchase Attitude, Subjective Norm, Motivations, Marketing Mix	4		✓		✓	
Hamzaoui-Essoussi & Linton (2014)	Estimated WTP[1]	Product Category, Perceived Risk, Brand Name	3		✓		✓	
Khor & Hazen (2017)	Purchase Intention	Attitude Subjective, Norm Perceived, Behavioural Control	3		✓		✓	
Hamzaoui Essoussi & Linton (2010)	Estimated WTP	Product Category, Perceived Functional Risk	2		✓		✓	
Wang & Hazen (2016)	Purchase Intention	Cost, Quality, Green Attributes, Perceived Risk, Perceived Value	4		✓		✓	
Abbey, Meloy, Guide, & Atalay (2015)	Attractiveness	Brand Equity, Negative Attributes, Quality Attributes, Green Attributes, Consumer Greenness, Price Discounts	6		✓		✓	
Guide Jr. & Li (2010)	Actual WTP	Bid ID, Bid Amount, Bid Time	3		✓			✓
Jakowczyk, Frota Neto, Gibson, & Van Wassenhove (2017)	Search Intensity; Listings Number	Search Intensity for New & Used Products; Listings Number For New & Used Products; Elapsed Time, New Price, Moving Parts, Weight, Personal Hygiene	9		✓			✓
Subramanian & Subramanyam (2012)	Actual WTP (Price Differential Between New and	Remanufacturer Identity, Seller Type, Seller Incumbency, Warranty Strength, Quantity Available, Demand Proxy, Selling	10		✓			✓

	Remanufactured Product)	Fee, Listing Time, Positive Feedback, Negative Feedback						
Pang, Casalin, Papagiannidis, Muyldermans, & Tse (2015)	Actual WTP (Price Differential)	Demand Proxy, Supply Proxy, Warranty Length, Positive Feedback, Negative Feedback, Seller Identity, Duration, Listing Time, Seller Incumbency	9		✓			✓
Frota Neto, Bloemhof, & Corbett (2016)	Actual WTP (Selling Price)	Product Description, Pictures, Product Condition, Positive Feedback, Negative Feedback, Memory, Product Generation, Shipping Cost, Warranty	9		✓			✓
Xu, Zeng, & He (2017)	Actual WTP (Selling Price)	Production Condition, Pictures, Negative Feedback, Seller Tenure, Autopay, Expedited Shipping, Return	7		✓			✓
This chapter	Actual WTP (Sales Level)	Overall Rating, Positive Feedback, Negative Feedback, Review Volume, Review Sentiments, Number of Helpfulness Votes, Number of Answered Questions, Product Description, Pictures, Discount Rate, Stock Availability, Warranty Information, Price Difference	13	✓	✓	✓		✓

[1] WTP: Willingness-to-pay

As Table 4.1 shows, most of the insights into consumer behaviour towards remanufactured products have been gained through the collection of primary data, in order to examine the key determinants that underpin customers' purchase intention and estimated willingness to pay (WTP) (Abbey, Meloy, et al., 2015; Hamzaoui-Essoussi and Linton, 2014; Hamzaoui Essoussi and Linton, 2010; Jiménez-Parra et al., 2014; Khor and Hazen, 2017; Wang and Hazen, 2016). The authors of these studies believe that a consumer will follow his/her purchase intention when making an actual purchase decision. However, in practice, a consumer's purchase behaviour does not always follow his/her purchase intention. For example, green awareness may be a driving force behind a consumer's intention to purchase a remanufactured product, but the actual purchase decision is based on the perceived quality and risk associated with the product, rather than on their initial intention (Khor and Hazen, 2017). To overcome this issue, another new research stream is initiated by using secondary data of real transactions to reflect real WTP, thereby exploring actual consumer purchase behaviour (Frota Neto et al., 2016; Subramanian and Subramanyam, 2012; Xu et al., 2017).

However, a major drawback of the previous studies is that they study consumer behaviour using a linear regression (LR) model which fails to capture the non-linear relationships between various predictors. As a result, its prediction accuracy deteriorates substantially when dealing with real-world datasets characterised by highly complex and non-linear relationships between multiple market factors.

4.1.3 Review on ML approaches for regression problems

In the BD era, ML approaches are increasingly used to overcome the drawbacks of LR and to generate data-driven decision-making. Empirical evidence suggests that companies that use data-driven decision-making can see significant improvements in both productivity and profitability (Bohanec et al., 2017). The key difference between the statistics-based LR and ML models is that the former uses a parametric approach in which the model structure is predetermined and the input–output relationship is forced to fit certain simplified assumptions. In contrast, the latter does not start from the model structure, but uses algorithms to determine input–output relationships from the dataset. Therefore, ML can approximate the complex and unknown non-linearity of noisy, high-dimensional datasets better than statistical methods.

Demand forecasting is naturally considered to be a regression problem in ML, and aims to accurately estimate the demand level of a product based on its relationships with a given set of independent input variables (i.e. predictors). Some examples of ML algorithms for regression problems include support vector machine (SVM), classification and regression tree (CART), artificial neural network (ANN), and K-nearest neighbour (KNN) algorithms. A recent review of big-data applications in supply-chain management (SCM) by Nguyen, Zhou, Spiegler, Ieromonachou, & Lin (2018) shows that ML has been widely adopted in traditional SCM to improve demand prediction. However, in CLSCs, the application of ML is largely underexploited and merely used for product return forecasting (see, e.g. Mazhar, Kara, & Kaebernick, 2007). The use of ML for remanufactured product demand forecasting has hardly been studied.

Despite the high level of predictive performance, ML models face resistance from users, as there is a perceived lack of comprehensibility (also known as interpretability) which requires practitioners to understand the insights behind the prediction of the model. In some areas, such as credit risk analysis and medical diagnosis, comprehensibility is even more vital than prediction accuracy (Martens et al., 2007).

While many ML models, such as the SVM, ANN, KNN and ensemble models, are widely criticised as being ‘black box’, the regression-tree approach is a type of ML that has gained popularity for creating a good balance between predictive performance and interpretation (De Caigny et al., 2018; Masci et al., 2018; Yang, Liu, et al., 2017). In single regression-tree models, such as CART and the M5 model tree, high interpretability derives from the tree-based graph in which one can quickly detect the most important variables used for node splitting and their variable interactions.

However, in more advanced regression-tree models, such as the ensemble-based random forest (RF) model, the tree visualisation is no longer possible. Alternative explanatory methods are needed to reveal the insights of such black box models. Two techniques which tackle this task effectively are variable importance ranking (VIR) and partial dependence plot (PDP). These methods are commonly used in many domains, such as education (Masci et al., 2018), ecology (Cutler et al., 2007), business risk management (De Bock, 2017), and supply-chain finance risk prediction (Zhu et al., 2019). To the best of the knowledge, however, they have not yet been applied in remanufacturing and CLSC management.

4.1.4Research objective and contribution

The literature review makes clear that it is necessary to use more sophisticated ML techniques to improve the accuracy of remanufactured product demand prediction (RPDP), which is essential for developing a marketing strategy for remanufactured products. This research therefore develops a data-mining approach that pursues two main objectives: (1) to obtain the most accurate and robust RPDP as possible by using the machine-learning approach; and (2) to analyse the RPDP by using VIR and PDP in order to gain an in-depth understanding of the online purchasing behaviours of consumers of remanufactured products, leading to the development of a practicable marketing strategy. A real-world database consisting of 5,693 remanufactured product listings on www.amazon.com is used to pursue the above objectives.

This research is one of pioneering studies which apply an ML approach to the demand prediction of remanufactured products. Its contribution is twofold: theoretical and practical. In theoretical terms, the research (1) develops a structured business analytics approach that can balance the trade-off between prediction accuracy and comprehensibility, thereby stimulating the use of black-box ML models; (2) provides a highly accurate and robust demand prediction for remanufactured products (a research area which remains largely understudied); and (3) sheds light on the non-linear behaviours of online market factors on remanufactured product demand. In practical terms, the research provides guidelines with which managers can develop effective online marketing and selling strategies for remanufactured products, thereby increasing the viability and profitability of remanufacturing/CLSCs.

4.2 Methodology

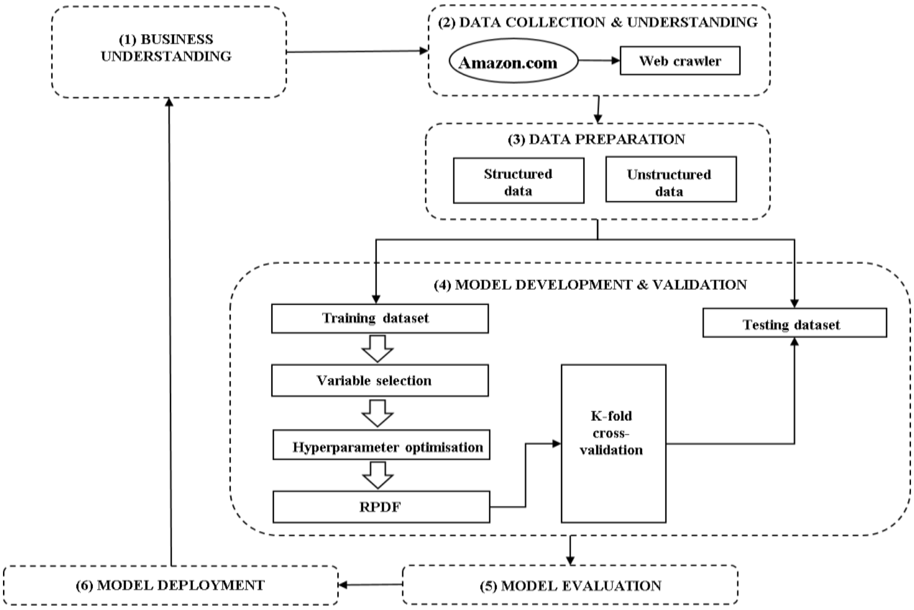


Figure 4.1 - Methodology framework

(Source: The author)

This research follows the Cross Industry Standard Process and Data Mining (CRISP-DM) framework, one of the most popular methodologies for data analytics (Oztekin et al., 2016). In line with the CRISP-DM, the data-mining approach in this research consists of six fundamental steps as shown in Figure 4.1: (1) Business understanding refers to the conversion of the business objective, RPDP in this research, into a data mining problem; (2) Data understanding identifies the data source and obtains the variables related to the problem; (3) Data preparation uses several data cleaning and transforming techniques to produce a well-structured dataset prior to analysis; (4) Predictive modelling includes variable selection, model development, hyperparameter tuning and validation; (5) Model evaluation measures and compares the predictive performance of the models based on different predefined error measurements; (6) Model deployment generates insights to assist managerial decision-making.

4.3 Data collection and use of variables

4.3.1 Research context and data collection

Consistent with previous studies of the online marketplace, www.amazon.com has been chosen as the data source for this study. Amazon classifies product conditions into three major categories — ‘new’, ‘certified refurbished’ (another term for remanufactured products), and ‘used’. To distinguish between remanufactured products and used products, Amazon defines remanufactured products as products that have been tested and certified by original equipment manufacturers (OEMs) or by qualified/specialised third-party remanufacturers to ensure their ‘like-new’ working condition. In this research, the author focuses on remanufactured products, as they are among the core products in CLSCs and have a direct impact on the manufacturing processes of OEMs (Zhou and Disney, 2006).

Amazon does not publicly reveal actual sales transactions. Instead, it uses the sales ranks to indicate the sales performance, which is recognised as a well-established proxy of customer demand (Archak et al., 2011; Dekkers, 2011; Hu et al., 2014; Li, Ch’ng, et al., 2016b). In general, the smaller the value of a product’s sales rank, the higher its customer demand. As there are scale effects in the data, the sales rank is estimated by its natural logarithm transformation (ln) rather than its level, as suggested by Chevalier & Mayzlin (2006). Hence, the dependent variable used in this study is expressed as follows:

$$customer.demand = \frac{1}{ln\ (Salesrank\ )} \tag{4.1}$$

where $Salesrank$ is Amazon's sales rank of the remanufactured product listing; and $customer.demand$ is the demand level of the product and also the dependent variable in this model.

The author develops a python programme to crawl the daily historical data of remanufactured products listed on www.amazon.com. For each product listing, the crawler was coded to capture a complete set of the data publicly available on the product page. It is noteworthy that Amazon sales ranks are frequently updated so that they reflect recent sales (with a maximum period of one months) (Chevalier and Mayzlin, 2006). Hence, in order to eliminate the potential simultaneity issue between independent and dependent variables, the author follows the approach of Chevalier & Mayzlin (2006), which predicts the sales rank at time *t*, based on lagged explanatory variables up to one month before time *t*. In particular, sales ranks were recorded on 30 May 2018, while all the predictors were recorded during the period up to 30 April 2018 (i.e. one month before the 30 May sales rank).

Our dataset includes remanufactured products that have comparable sales ranks in the Amazon Electronics category. where remanufacturing activities are particularly important, because of the end-of-life environment effect of products (Subramanian and Subramanyam, 2012). This category includes technological products such as cell phones, computers, cameras and GPS navigation.

4.3.2 Theoretical background and variable description

4.3.2.1 Theoretical background

In the contract and economic theory, the term 'information asymmetry' refers to market interaction under conditions in which sellers have more or better information about the quality of the product and service than buyers (Boulding and Kirmani, 1993). Such an information gap would increase the buyer's quality uncertainty and the perceived risks of buying a low-quality product, making buyers willing to pay no more than the average market price. Consequently, sellers have no incentive to sell high quality products, and are likely instead to keep low quality products on hand (Frota Neto et al., 2016). This phenomenon was first described by Akerlof

(1970) as the ‘market of lemons’. To prevent the emergence of lemon markets, sellers often apply market signal theory, in order to address the information asymmetries of customers (Boulding and Kirmani, 1993; Wells et al., 2011). The theory provides a framework with which to understand how sellers can use extrinsic cues (i.e. signals) to convey information about their product and service quality to buyers, in order to reduce their perceived uncertainty and to facilitate trade in an environment featuring high information asymmetries (Li et al., 2015).

As it focuses on the online marketplace of remanufactured products, the theoretical foundation of this study is primarily based on market signal theory. This is because: (1) empirical research has suggested that the burden of information asymmetry between sellers and buyers is exaggerated in the e-marketplace, as buyers have no physical interaction with the products prior to their purchase (Li et al., 2015); and (2) the information asymmetry of product quality seems to be more severe for remanufactured products than new products, due to the higher customer uncertainty and perceived risks associated with unobservable remanufacturing processes (Tereyağoğlu, 2016).These factors increase the need of customers for additional information about the product, which ultimately means that quality cues have a stronger influence on customers’ purchasing decision-making (Frota Neto et al., 2016).

Seller-generated content (SGC) and user-generated content (UGC) can both be signalling factors. Following previous research (e.g., Frota Neto et al., 2016; Subramanian & Subramanyam, 2012; Xu et al., 2017), this study uses a number of the SGC and UGC variables which are available on Amazon to predict demand for the remanufactured product. The description and measurement of each variable are presented as follow.

4.3.2.2 Description and measurement of variables

- **Positivity of product description**

A common way for sellers to ease the problem of information asymmetry with customers is to use textual descriptions of products. The seller-provided product description is typically written in natural language with structured formats. It serves as a snapshot which summarises the key selling points that favourably differentiate their product from others. Sellers of remanufactured products often use keywords that positively describe the condition and quality of these products, such as ‘like new condition’, ‘certified refurbishment’, ‘testified’, ‘mint’, ‘excellent condition’ and ‘no scratches’. The effect of such positive keywords on customer WTP has been previously studied, with mixed findings. For example, van Heijst, Potharst, & van Wezel (2008) find that positivity of product description is the most influential predictor of customer WTP to remanufactured products, compared to the number of product pictures and seller feedback ratings. In contrast, Frota Neto et al. (2016) find no significant statistical evidence for such a relationship. Despite the mixed result, it is therefore still very likely that customers would use such condition-related keywords as a cue about remanufactured product quality. As such, based on market signalling theory, the author expects that the positivity of product description will have a significant, positive effect on customer demand for remanufactured products. The technical detail of how this variable is constructed is described as follows.

After being extracted from the Amazon product page, each individual product description is transformed into a vector space model using three steps for the unstructured data preparation (i.e. tokenisation, stopword filtering and part of speech (POS) tagging), as explained above. Each dimension of the vector corresponds to a separate term which appears in the product descriptions. The author defines the terms using both bag-of-word (i.e. single word) and bigram (i.e. phrase of two words), to avoid losing the text semantics. The occurrence frequency of terms is then counted to build up the dictionary. The author applies a filter condition that removes terms with an occurrence frequency lower or equal to 1 in order to decrease the computational time and to exclude misspellings and infrequent words, as suggested by van Heijst, Potharst, & van Wezel (2008). The size of the dictionary obtained is significantly reduced after applying this filter. Next, the author follows the approach of Frota Neto et al. (2016). The author asks two coders, who are not the co-authors of this research, to independently select from the given dictionary a list of keywords that positively indicate the high quality or performance level of the remanufactured product and its remanufacturing process, such as “like new”, “excellent condition”, “no scratch”, “tested”, “cleaning”, etc. Both lists are then checked for consistency by each of the authors of this research. After cross-checking, the final list is consolidated. Finally, the positivity of product description for each remanufactured product is measured by counting the number of times the positive keywords appear in the description. The author observes a good distribution in the value range of this variable, which suggests that the obtained list has captured most of the commonly used positive keywords. Examples of some common positive keywords from the list are presented in Table 4.2.

Table 4.2 - Examples of the positivity of product description

Product description	**Positivity of product description**
This refurbished product is **tested** and **certified** to look and work **like new**. The refurbishing process includes **functionality testing**, basic **cleaning** and **repackaging**. A minimum 90-day **warranty**, and may arrive in a generic box. Only select sellers who maintain a **high performance** bar may offer **certified** refurbished products on Amazon.com.	9
Renewed products look and work **like new**. These pre-owned products have been **inspected** and **tested** by Amazon-**qualified** suppliers, which typically perform a full **diagnostic test**, and a **thorough cleaning** process.	6
This computer has been **certified** refurbished by a Microsoft **authorized refurbisher**, ensuring the **highest** levels of quality and support.	3

- **Number of product pictures**

In addition to textual quality cues, sellers also use product pictures as visual quality cues because some product attributes are difficult to describe in words, such as signs of wear on shoes (Frota Neto et al., 2016; van Heijst et al., 2008). Therefore, it is expected that the number of product pictures provided by the seller will have a positive effect on the remanufactured product demand.

- **Warranty information disclosure**

Due to the monetary and legal risks involved, the signalling role of warranties as a quality cue has gained much more attention than textual and visual cues for both new products (Li et al., 2015) and remanufactured products (Alqahtani and Gupta, 2017; Subramanian and Subramanyam, 2012). Although warranties may subsequently incur extra costs for sellers/manufacturers, they can bring additional profits at the point of sale by reducing customers’ concerns about product reliability during their purchase decision-making. This positive effect of warranties on online sales is found to be significant for new products (Li et al., 2015), but has not yet been confirmed for remanufactured products (Abbey, Meloy, et al., 2015; Atasu et al., 2010; Jiménez-Parra et al., 2014; Subramanian and Subramanyam, 2012).

Instead of focusing on warranty strength (i.e. the duration of warranty), this study examines a new angle that has not been yet gained much research attention — the ways in which warranty information disclosure affects customer purchasing behaviour towards remanufactured products. According to Tereyağoğlu (2016), sellers often see the disclosure of full information about the warranty and entire remanufacturing process as one of the strategies which can reduce customers' uncertainty about the remanufactured product quality, thereby further increasing sales. This strategy is in line with transactional cost theory and the linkage principle of Milgrom & Weber (1982), suggesting that voluntary information disclosure by sellers can help reduce information asymmetries and save customers the cost of acquiring information, thereby increasing their WTP.

Some Amazon sellers of remanufactured products provide technical documents which typically specify the details of the remanufacturing process, product condition and the full terms and conditions of the product warranty. Other sellers require customers to take extra time and effort, such as by contacting the seller by email or being directed to other websites for further warranty details. To reflect the impact of warranty information disclosure, the author creates a binary variable, which is equal to '1' if the Amazon seller provides a technical document specifying the full warranty terms and conditions in detail, but equal to '0' if no such document is given. Following the above discussion, the author would expect that remanufactured products with full warranty information disclosure are associated with high customer demand, whereas those without such information are associated with lower demand.

- **Stock information**

Stock information is another binary variable which is considered to be a potential predictor of remanufactured demand. This variable is equal to '1' if the seller specifies a limited quantity of stock left; otherwise, it is equal to '0'. The author adopts this variable from the marketing perspective. By leveraging the psychological effect of the scarcity principle, sellers use the limited availability of a product to increase its perceived quality and boost sales (Cialdini, 2009). The existing literature contains abundant empirical evidence on the importance of the scarcity effect for purchasing decisions (Swami and Khairnar, 2006). However, it is not known whether this effect also exists for remanufactured product. In this study, the author expects that the scarcity principle still applies. As such, remanufactured products with limited availability will be associated with a higher demand level than those with abundant stocks.

- **Price difference between remanufactured and corresponding new products**

As Abbey, Blackburn, et al. (2015) state, managers often experience fear and uncertainty when making pricing decisions about remanufactured products — fear of the risks of cannibalisation by new products and uncertainty about customers' WTP for remanufactured products. Following the law of demand in microeconomic theory, which holds that a lower price will lead to a higher demand, many firms have been setting the prices of their remanufactured products 10% to 80% lower than those of the corresponding new products (Ovchinnikov, 2011). In this way, the seller uses price as a marketing tool with a discounting effect. This practice is supported by a small number of studies which found that price discounts should have a positive, linear effect on the perceived attractiveness and sales of the remanufactured product (Abbey, Blackburn, et al., 2015). However, the existing literature has also found some evidence for the non-linearity or even negative effects of price discounts on the attractiveness of, and demand for, remanufactured products (Ovchinnikov, 2011). This study therefore aims to determine whether a price differential between a remanufactured product and the equivalent new product will increase demand for the remanufactured one.

The variable representing the price difference is expressed in Eq (4.2):

$$Price\_difference\ = \frac{Average\ selling\ price\ of\ new - selling\ price\ of\ remanufactured}{Average\ selling\ price\ of\ new} \times 100\% \tag{4.2}$$

The method used to obtain the price of corresponding new products from Amazon is described as follows.

On most of the product pages, Amazon has a 'Compare with similar items' function, providing a list of the most relevant products to support each customer's purchase decision-making process by reducing their product screening and evaluation cost (Zhang et al., 2018). Each product in the list is presented with key information such as its price, title and image, a link to its main listing page, its customer ratings and shipping details. The price differential between each remanufactured product and its new counterpart (Eq.4.2) is obtained by taking the average price of all the equivalent new products in its recommendation list. The average price is used because it is a more robust reference price than the lowest price, as Subramanian & Subramanyam (2012) suggested.

- **Product promotion rate**

As well as selling remanufactured products at a lower price than the equivalent new products, sellers also offer some promotional discounts, aiming for an immediate short-term increase in sales. According to transactional utility theory, higher discount rates increase customers' utility, and therefore lead to higher sales (Dekkers, 2011). In this research, the author expects there is a positive link between a promotional discount rate and remanufactured product demand. The discount rate is collected directly from each Amazon product page.

- **Overall product rating**

Like most e-commerce sites, Amazon provides an overall product rating which is the average of the ratings from previous customers. This numerical index ranges from the lowest rating of 1 to the highest of 5. It indicates the valence dimension of the seller's reputation and the overall attractiveness to consumers based on their past sales performances and future prospects, which provides signals that reduce the burden of information asymmetry and build customer trust in the sellers (Li et al., 2015). Previous studies have repeatedly found that the overall product rating has a positive and linear effect on sales of new products (Archak et al., 2011; Chevalier and Mayzlin, 2006; Li, Ch'ng, et al., 2016b; Li et al., 2015). However, the effect of the product rating for remanufactured products has not been sufficiently studied, and this study hopes to address that gap. The author expects the overall product rating to have a positive effect on remanufactured product sales.

- **Number of service failures and number of service successes**

Previous studies have found that the effect of negative customer feedback on sales of new products can be more pronounced than that of positive feedback, as the damage to the seller's reputation significantly increases the risk perceived by prospective customers (Chevalier and Mayzlin, 2006; Cui et al., 2012; De Maeyer, 2012). Subramanian & Subramanyam (2012) confirm the finding for remanufactured products. However, their finding is based on the linear model, which motivates us to re-examine whether such negative bias remains significant in the non-linear model. As a result, this study aims to determine whether the number of service failures and the number of service successes are important predictors of remanufactured product demand, and whether the effect of service failures on sales is more significant than that of service successes. The method of obtaining these two variables from Amazon is explained as follows.

Together with the total number of cases of customer feedback, Amazon also provides the distribution in percentage of customer ratings of 1, 2, 3, 4 and 5 stars on each product listing page. The author collects these metrics. Referring to prior research (e.g., Frota Neto et al., 2016; Xu et al., 2017), the total number of neutral (3 stars) and negative (1 and 2 stars) customer ratings can be used to measure the seller's service failures. Likewise, the total number of positive (4 and 5 stars) customer ratings are used to measure the seller's service successes. It is noted that this study uses the absolute number of feedback counts rather than the percentages, to avoid misrepresenting the seller reputation, as suggested by Xu et al. (2017) and by Subramanian & Subramanyam (2012).

- **Number of helpfulness votes**

For each product, the author takes the sum of helpful votes received from all customer reviews. This metric can serve as a quality indicator of the reviews, according to other consumers. A large number of helpful votes indicates that the product reviews are of high quality, meaning that the reviews contain a large amount of helpful information and can influence the purchasing decisions of other buyers. The positive effect linking the helpfulness of reviews and the online sales of new products have been found in previous literature (e.g. B. Li et al., 2016). In particular, this effect can become more powerful for less popular products (De Maeyer, 2012). This study aims to test the validity of these findings in the case of remanufactured products.

- **Number of answered questions**

As well as indirect interaction between users through helpfulness votes, Amazon has initiated ‘Customer questions and answers’, which allows customers to directly ask and respond to each other’s questions. According to trust transference theory, a large number of answered questions can stimulate social interactions which further boost online trust for higher purchasing intention (Ng, 2013). Previous studies have found a significant and positive effect of this variable on the sales predictions for new products (Dekkers, 2011; Li, Ch’ng, et al., 2016b). Hence, it is expected that the variable will have a strong predictive power on demand for remanufactured products.

- **Average sentiment of customer reviews**

In addition to the numerical ratings of the product, Amazon also allows a customer to post a textual review that provides a context-specific explanation of his/her shopping experience with varying degrees of polarity sentiment (e.g. strongly or moderately negative, positive or neutral). Such polarity sentiments provide rich information to the reader who goes beyond numerical ratings (Hu et al., 2014). Although people often consider the sentiment of a textual review to be consistent with the numeric product rating, previous research indicates that these two are not always aligned (Hu et al., 2014). Archak et al. (2011) posit that only using numerical ratings may not fully capture the effect of product reviews on customer purchasing behaviour, unless the reader of a review has exactly the same preferences as its writer. According to trust transference theory, when potential customers read online reviews, their emotions may be affected by the sentiments expressed in the product review, and their emotional status may influence their evaluation of the product and direct their purchase decisions. Previous studies have found that review sentiments play a significant role in the prediction of new product demand (Archak et al., 2011; Dekkers, 2011; Hu et al., 2014). In this study, the author examines whether such strong predictive power is also true of remanufactured products. The author expects that remanufactured products indicating positive review sentiment would be associated with high customer demand. The technical detail of the construction of the variable representing the review sentiment of each remanufactured product is explained as follows.

The author collects all the textual reviews of each remanufactured product. Each review is then transformed into numerical data using the three text-processing steps: tokenisation, stopwords filtering and bagging of words, as used to construct "*Positivity of product description*" variable above**.** After the preparation process, the textual data is in a structured format of vector space and ready for sentiment analysis.

Sentiment analysis is a research technique that systematically analyses written and spoken content either through computers or through a qualitative method, in order to extract, detect and analyse customers' opinions, emotions, attitudes and subjectivities in texts (Xiang et al., 2015). One main objective of sentiment analysis is to identify the polarity of the review (i.e. positive, negative or neutral) and also the polarity strength (i.e. strongly or moderately) (He et al., 2015).

In this study, SentiStrength is used to conduct sentiment analysis. Product reviews are usually short and informal, and SentiStrength has been proved to be useful and accurate when analysing short and informal messages (Stieglitz and Dang-Xuan, 2013). Moreover, many sentiment analysis tools can only measure sentiment polarity (i.e., positive, neutral, negative), while SentiStrength is also capable of measuring sentiment strength. SentiStrength refers to a lexicon of emotional words and terms and has certain linguistic rules for issues including spelling corrections, negations (e.g., "not good"), booster words (e.g., "very sad"), amplifications (e.g., "haaaaaaaappy"), word weighting and emoticons (Stieglitz and Dang-Xuan, 2013). Each word (referred to as a 'term') contained in the product review is classified for polarity and strength. Polarity indicates whether the word is positive or negative, and strength indicates whether the word sentiment is strong or weak. For example, "ache" is identified as -2, which means mildly negative, "excruciating" is identified as -5, which means strongly negative, and "great" is identified as 3, which means moderately positive. For each review text, the positive and negative sentiment scores are measured separately, where the positive sentiment score ranges from 1 (no positive sentiment) to 5 (strong positive sentiment), and the negative sentiment score ranges from -1 (no negative sentiment) to -5 (strong negative sentiment). The overall sentiment strength of each review text is then calculated as the sum of the positive sentiment score and the negative sentiment score, ranging from -4 to 4. For example: the review sentence, "The refurbished camera works *perfectly* [+3]", has a positive sentiment score of 3 and a negative sentiment score of -1. Therefore, the overall sentiment strength of the sentence is 2, i.e. 3 + (-1). For each product, the author calculates the overall sentiment strength for each review. The final sentiment strength of the product (used as a predictor in the sales prediction model) is then measured by taking the average of all the overall sentiment scores of its reviews.

- **Brand equity**

Brand equity can be used to reduce customers' perceived risks and uncertainties about product quality and increase their perceived level of trust. When a product has high brand equity, the brand is considered to have high value, and the product is associated with high levels of quality, reliability and awareness. Hence, there is likely to be high demand for it. Previous research has identified brand equity to be important for remanufactured products (Abbey, Meloy, et al., 2015). Therefore, in this study, brand equity is included in the model as a control variable. More specifically, the author classifies brand equity as high or low, based on the customer voting on www.ranker.com, where top brands are listed. This website provides the brand rank lists for a wide range of niche markets and product types. These rank lists are voted on by millions of visitors every month, and therefore accurately reveal the brand conception of the crowd. If a product brand is listed on www.ranker.com, it is marked as being of high brand equity; if not, it is marked as being of low brand equity.

4.4 Data preparation

There are both structured (e.g. numerical data) and unstructured (e.g. textual data) types of variables in the studied Amazon dataset that require different pre-processing techniques before they can be analysed.

4.4.1Structured data preparation

The main task here is handling the missing data in the dataset. Previous studies (e.g. De Caigny et al., 2018; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012) have suggested that for continuous variables with more than 5% of missing data, the missing values should be treated using appropriate imputation procedures. However, for those with less than 5% of missing values, the missing values should be removed to limit the effect of imputations (Little, 1988).

In addition, all the categorical variables should be transformed into dummy variables (1 or 0), which are then treated as numerical variables.

Outlier treatment is also carried out, to ensure data consistency. Outliers are out-of-ordinary values that are typically defined as being more than three standard deviations (SD) away from the mean. Univariate outliers among continuous variables are visually detected using a box-and-whisker plot, and winsorised into acceptable values that are within 1.5*IQR (Interquartile ranges), using R programming.

To address the scale effect between continuous independent variables, their values are normalised into the range between 0 and 1.

4.4.2 Unstructured textual data preparation

Customers' textual reviews and sellers' textual product descriptions can be highly unstructured and extremely noisy. Specific text pre-processing techniques are therefore required to transform such data into a usable, structured format. The techniques used in this research include: (1) tokenisation: breaking the text sentences into word vectors; (2) stopwords filtering: removing unnecessary stopwords for grammar rules (e.g. a, an, the); and (3) part-of-speech (POS) tagging: referring to a category of words having similar grammatical properties such as noun, verb, adjective, etc. The POS tagging is very helpful as it filters out all the words that do not convey sentiments (i.e. relevant or meaningful information). More details of these pre-processing steps can be found in Alaei, Becken, & Stantic (2017). As a result, each of the customer reviews and the seller's product descriptions are transformed into vector space models, in which each vector dimension corresponds to a separate term (or keyword) that contains sentiments. The term vectors were used to construct two of the predictors used in the proposed remanufactured product demand model, namely the positivity of product descriptions and average review sentiments, as described in Section 4.3.2.2.

4.5 Model development and validation

As Figure 4.1 shows, the author employs three specific procedures that are important for a fair and valid comparison of the performance between different ML models. These are variable selection, hyperparameter optimisation and k-fold cross-validation (CV).

4.5.1 Variable selection

After the well-structured dataset is prepared and aggregated, the variables are selected as predictors in the RPDP model. Variable selection is a critical step in ML applications, as using excessively large datasets with many irrelevant/noisy variables often slows the algorithm, consumes more resources and can even damage predictive performance. Based on the concept of variable relevance, Nilsson, Peña, Björkegren, & Tegnér (2007) classified variable selection into two categories of problem: (1) *all-relevant problems* — finding all strongly and weakly relevant variables; and (2) *minimal–optimal problems* — finding the subset of strongly relevant variables and removing the subset of weakly relevant variables that contains only redundant information. Variable relevance here is a qualitative measure and is independent of classifier types, which is distinct from the variable importance in the VIR analysis, as discussed later (Rudnicki et al., 2015).

To increase the applicability of the research to future work, the author tackles both of the feature selection problems. The Boruta algorithm is adopted for the all-relevant set, while the recursive feature elimination (RFE) algorithm is adopted to select the minimal–optimal set of predictors. All the predictive algorithms for RPDP discussed in Section 4.4.3.2 are run with these two input sets. The procedures of both algorithms are detailed as follows.

In the RFE algorithm, the author first fits the model with all the variables and rank the importance of each variable based on the model performance. Let S_1 > S_2 > ...> S_s be the ordered sequence that indicates the number of variables to retain. For each iteration, the most important variable of the subset S_i is retained, and the model is refit with the preformation re-evaluated. As a result, the author selects the smallest subset of the most important predictors, which can give the most comparable accuracy of demand forecasting.

The principle of the Boruta algorithm arises from an RF model that injects greater randomness into the system; the result is therefore quite robust to noise and is unbiased. The algorithm first extends the original training dataset with the duplication of all predictors. The extended dataset is shuffled, in order to reduce the correlation with the response variable. The importance of each variable is measured based on its Z-score, which indicates how many standard deviations (σ) from the mean (μ) a data point (x) is, as expressed by $z = \frac{x-\mu}{\sigma}$. At every iteration, the Z-score of each original variable is compared with the maximum Z-score of its randomised variable (the so-called ShadowMax). If the variable has a Z-score higher than its ShadowMax, then it is considered to be an important predictor. Variables with Z-scores significantly lower than their ShadowMax are assessed as unimportant and removed from the dataset. Tentative variables are those with Z-scores very close to their ShadowMax; they can either be removed or retained. More details of the Boruta algorithm can be found in Kursa & Rudnicki (2010).

4.5.2 Regression tree predictive model descriptions

Regarding the predictive algorithms for RPDP, three regression tree models (CART, M5 and RF) are employed.

- **CART model**

Like most decision tree algorithms, CART adopts a ‘divide and conquer’ strategy in order to construct the tree-based model. It aims to identify a predictor and its breaking-point value as the tree node for the binary splitting of the training space into the most homogeneous (or purest) subsets. The formal description of the CART model is presented as follows.

Let $y_m^{(p)} = (y_1^{(p)}, y_2^{(p)}, \dots, y_m^{(p)})$ and $X_n^{(p)} = (x_1^{(p)}, x_2^{(p)}, \dots, x_n^{(p)})$ be the set of m observations and n predictors at the current node p, respectively. Let us assume that $x_s^{(p)}$ is the predictor selected for splitting with its unique break-point value, θ. The observation data in the two child nodes resulting from the binary split in $x_s^{(p)}$ at the parent node p are $y_{Left}^{(p)}$ and $y_{Right}^{(p)}$, where $y_{Left}^{(p)}$ contains those data of $y_m^{(p)}$ that have corresponding values of $x_s^{(p)} \leq \theta$ and $y_{Right}^{(p)}$ contains those with $x_s^{(p)} > \theta$. The splitting criterion for choosing $x_s^{(p)}$ and θ is based on how much the split contributes to the reduction in loss function, $\Delta y_m^{(p)}$, which can be mathematically expressed as:

$$\Delta y_m^{(p)} = L_{MSE}(y_m^{(p)}) - \left\{\frac{m_L^{(p)}}{m} L_{MSE}(y_{Left}^{(p)}) + \frac{m_R^{(p)}}{m} L_{MSE}(y_{Right}^{(p)})\right\}, \quad (4.3)$$

where $m_L^{(p)}$ and $m_R^{(p)}$are the numbers of observation cases in $y_{Left}^{(p)}$ and $y_{Right}^{(p)}$, respectively. $L_{MSE}(.)$ is the loss function that is usually measured using the mean squared error (MSE) in regression trees:

$$L_{MSE}(y^m) = \sum_{i=1}^{m} \left(y_i^{(p)} - \hat{y}^{(p)}\right)^2, \quad (4.4)$$

where the predicted value $\hat{y}^{(p)}$ is typically measured as the mean of the observations in $y_m^{(p)}$.

In the CART model, there are two tuning parameters — the maximum tree depth to growth as the stopping criterion for tree growth and the threshold value of loss function reduction $\Delta y_m^{(p)}$ as the splitting criterion. The splitting process terminates when the $\Delta y_m^{(p)}$ varies slightly (e.g. less than 5%) or when there are only a few observations remaining (e.g. four or fewer). As experiments indicate that the performance is not very sensitive to the particular choices of the $\Delta y_m^{(p)}$ threshold, the author only includes the maximum tree depth in the hyperparameter tuning step (Witten et al., 2011).

Despite its high interpretability and efficiency, there are some practical issues affecting both ‘divide and conquer’ steps in the CART algorithm: (1) using the greedy search to grow the tree is an efficient approach, but it only returns local optimality, which makes the tree very sensitive to even a small change in a dataset, and bias can occur if the dataset is imbalanced; (2) the high variance of the tree leads to the instability of predicted outcomes, because the prediction rule based on the mean value is too simple (Witten et al., 2011). As a result, there have been a few attempts to improve these shortcomings of CART, such as the M5 model tree and RF.

- **M5 model tree**

Introduced by Quinlan (1992), the M5 model tree attempts to improve the CART prediction rule by applying an LR model to predict the value in each leaf node, rather than taking the mean value. The construction of the model is also based on the 'divide and conquer' approach. At the first stage, the tree is grown with the same principle as CART. The only difference is that CART uses standard deviation of the output as the loss function, which therefore maximises the expected standard deviation reduction (SDR). The details of the M5 model are presented as follows**.**

In the M5 model, the maximum of the expected standard deviation reduction (SDR) is used as a node-splitting criterion:

$$SDR = SD(T) - \sum \frac{|T_i|}{|T|} SD(T_i), \tag{4.5}$$

where T is a set of observations at the parent node, and T_i is the subset of observations that results from the node splitting based on the chosen predictor. For the pruning method to trim down the overgrown tree at the second stage, Quinlan (1992) suggests replacing the internal nodes of the tree with LR models, which results in a predicted value of the dependent variable. Alternatively, the predicted value can be smoothed to improve the prediction accuracy of the tree as a whole. A more detailed analysis of the M5 learning algorithm can be found in Quinlan (1992) and Witten, Frank, & Hall (2011). For hyperparameter tuning of M5, the author needs to decide whether or not a pruning and smoothing rule should be in place.

- **RF model**

Random forest (RF) adopts the bagging ensemble method which combines multiple base classifiers with equally distributed weights, in order to increase the predictive performance of a single decision tree (Witten et al., 2011). The detailed process of bagging in RF is described as follows.

The random forest (RF) uses a large number of de-correlated decision tree models to vote for the most influential prediction. Breiman (2001) defines the RF as an ensemble of tree-based classifiers $\{h(m_{try}, \Theta_k), k = 1, 2, \ldots\}$ where the $\{\Theta_k\}$ is a set of k vectors that are independent, evenly distributed and randomly drawn from the original training data. For each random vector Θ_k, a regression tree (i.e. the CART tree in this study) is grown to the maximum depth of J_{node} using m_{try} inputs selected from the entire input space at random. As a result, we have a forest of k CART-based trees from which a final prediction can be made via a majority vote. There are three tuning parameters in this model — the number of trees, k; the maximum depth, J_{node}; and the number of inputs selected at random, m_{try}. As suggested in previous studies, the ML pays particular attention to the tuning value of m_{try} while the model performance is not very sensitive to changes in the values of J_{node} and k (Krauss et al., 2017).

By injecting randomness into both subsampling and input selection, RF has several distinct advantages, including (1) high prediction accuracy; (2) relative robustness to data outliers and noise; (3) higher speed, compared to other bagging and boosting models; (4) unbiased (low variance) internal error estimations, predictive strength, correlation and variable importance; and (5) easy hyperparameter tuning (Breiman, 2001). Therefore, RF has been commonly used for complex forecasting problems in diverse areas, such as finance (Krauss et al., 2017), ecology (Cutler et al., 2007) and many others. Using the CART tree as the base classifier, it is, however, difficult to interpret the RF results when numerous random single trees are added to the forest model. Therefore, like most other powerful ML prediction models (e.g. ANN, SVN), RF shares the main disadvantage of being a ‘black box’ (Witten et al., 2011).

4.5.3Hyperparameter tuning

It is essential to make a comparison between different ML models based on the best performance of each one, obtained by running with their optimal parameter settings (De Caigny et al., 2018). Therefore, the author performs the sensitive analysis of each model with regard to its own tuning parameters and using both the Boruta and RFE sets of inputs.

4.5.4K-fold cross validation (CV)

When comparing the prediction accuracy of multiple models, it is common to use k-fold CV to avoid the bias issue relating to data sampling (Oztekin et al., 2016). Given the sample size used in this research, the author runs all models with ten-fold CV, because the empirical research shows that k=10 is the optimal number of folds that optimises the computational time while also minimising bias and variance issues during the validation process (Kohavi, 1995). To totally remove the resampling bias when dividing *k* fold, the author further repeats ten-fold CV three times.

4.6 Model evaluation and deployment

4.6.1 Model evaluation

Three commonly used statistical measurements — mean absolute error (MAE), root mean square error (RMSE) and coefficients of determination (R^2) — are calculated in order to evaluate the predictive performance of different ML models. The higher the value of R^2, the better the predictive performance achieved by the model. In contrast, the lower the value of MAE and RMSE, the better the predictive performance of the model.

4.6.2 Model deployment

The author performs two analyses, VIR and PDP, in order to interpret the predictive power and marginal effect of the predictors of RPDP.

4.6.2.1 Variable importance ranking (VIR)

A simple way to extract insights from ML-based prediction is to rank the importance of the explanatory variable, based on its predictive strength to the response variable. In this study, the VIR is performed using the method of information fusion-based sensitivity analysis (IFSA) which is commonly used for the VIR in data mining and ML literature (e.g. Chen, Chen, & Oztekin, 2017; Dag, Oztekin, Yucel, Bulur, & Megahed, 2017; Delen, Oztekin, & Tomak, 2012; Oztekin, 2016, 2018; Oztekin, Delen, & Kong, 2009; Oztekin, Delen, Turkyilmaz, & Zaim, 2013; Sevim, Oztekin, Bali, Gumus, & Guresen, 2014).

Sensitivity analysis (SA) is used to measure the relative importance of each independent variable in the prediction model by measuring how the error function would change when a specific variable is excluded from the input set (Oztekin, 2018). The more sensitive the model is to the inclusion/exclusion of a specific variable, the higher importance that variable has to the prediction of the output variable. This method is often used after the ML model is trained to rank the importance of each variable based on its sensitivity score defined in Eq. (4.6) (Delen et al., 2012):

$$S_i = \frac{Var(\mathbb{E}(F_t|X_i))}{Var(F_t)} \tag{4.6}$$

where S_i is the sensitivity score of the independent variable i in the model. $Var(F_t)$ is the unconditional output variance. In the numerator, the expectation operator $\mathbb{E}$ is first used to call for an integral over X_{-i} (i.e. over all input variables except X_i), and then the variance operator Var is applied to call a further integral over X_i. The importance of a particular variable is then calculated as the normalised sensitivity (Delen et al., 2012).

Nonetheless, it is expected that the SA results of each prediction model comes out to be somewhat different from each other. Therefore, to combine the SA results for higher robustness and lower bias in VIR, the information fusion (IF) technique is used (Oztekin, 2018). Together with Eq (4.6) above, the sensitivity score of the variable m with information fused by n prediction models can be computed by Eq. (4.7) (Sevim et al., 2014):

$$S_{m(fused)} = \sum_{i=1}^{N} \omega_i S_{im} = \omega_1 S_{1m} + \omega_2 S_{2m} + \cdots + \omega_n S_{nm} \tag{4.7}$$

where $\omega_1, \omega_2, \ldots, \omega_n$ refer to the weighting coefficients of each individual prediction model, namely, $f_1(x)$, $f_2(x)$, …, $f_n(x)$, respectively. The weights are normalised (i.e. $\sum_{i=1}^{N} \omega_i = 1$) and assigned proportionally to the prediction performance of each model. In other words, the greater the accuracy of the prediction model, the greater the weight is assigned to that model (Sevim et al., 2014).

However, the major limitation of VIR analysis is that it can only tell managers which predictors are important and which are not important, but cannot explain why. Therefore, despite its great popularity in practice, it is commonly criticised for being too theoretical and unable to provide a deep understanding of the subject matter (Grömping, 2015). To overcome this limitation, the author employs the PDP method to extract more insights from RPDP, as discussed in the following section.

4.6.2.2 Partial dependence plot (PDP)

In practice, managers will often want to know not only which predictors are most important, but also how they affect the response variable. For ML models, especially black-box ones such as RF, a graphical visualisation such as a PDP is one of the most effective ways of determining this. PDP can visualise the non-linear, complex marginal effect of single and multiple (usually two) predictors on the response variable, while also taking into account the average effects of other predictors. Two points worth noting about using a PDP are that: (1) a PDP can only partially provide the marginal effect of a variable over a certain range of its values rather than its complete behaviour; and (2) a PDP can be misrepresented in the presence of a high-order interaction or a strong correlation between predictors (Goldstein et al., 2015). A brief description of Friedman's PDP is described as follows.

Given a dataset D containing N observations y_k of the response variable y for $k = (1, 2, \ldots, N)$ and P predictors indexed x_k^i for $i = (1, 2, \ldots, P)$ and $k = (1, 2, \ldots, N)$, the model produces the predicted values $\hat{y}_k$ of y in the form:

$$\hat{y}_k = F\left(x_k^1, x_k^2, \ldots, x_k^p\right) \tag{4.8}$$

where $F(\ldots)$ represents some mathematical expressions. In the case of a single predictor x_j, Friedman's PDP is computed using the following average function and plotted over the observed range of x value (Masci et al., 2018):

$$\Phi_j(x) = \frac{1}{N}\sum_{k=1}^{N} F(x_k^1, \dots, x_k^{j-1}, x, x_k^{j+1}, \dots, x_k^p) \tag{4.9}$$

The function $\Phi_j(x)$ implies how the change in the value of the predictor x_j affects the model predictions $\hat{y}_k$ while 'averaging out' the effect of all other predictors.

Likewise, the dual effect of the two predictors x_i and x_j on the response variable y can also be defined analogously by the following bivariate partial dependency function (Masci et al., 2018):

$$\Phi_{ij}(x,y) = \frac{1}{N}\sum_{k=1}^{N} F\left(x_k^1, \dots, x_k^{i-1}, x, x_k^{i+1}, \dots, x_k^{j-1}, y, x_k^{j+1}, \dots, x_k^p\right) \tag{4.10}$$

This function is still fairly easy to visualise and interpret using a contour plot. However, Friedman's PDP loses its comprehensibility advantage when describing the interaction effect between three or more predictors. The interaction effect between more than two variables is therefore beyond the scope of this research.

4.7 Results and discussions

Following the data collection described in Section 4.3, the author obtains a database with 5,693 remanufactured products listed on www.amazon.com. These include 169 remanufactured products which have no corresponding new products for price comparison. Since the missing data comprises less than 5% of the dataset, the author removes these products, as explained in section 4.4.1. In addition, there are also 3,957 remanufactured products that have no customer review data. Given the fact that five out of the thirteen independent variables in this model are consumer-generated data, any imputation approach used to handle the missing data of customer reviews may mislead the research findings. The author therefore excludes these products from the experiment. As a result of the pre-processing step, the author has a database of 1,567 remanufactured products in total, along with 5,512 product pictures and 201,514 customer reviews. The dataset is split into two subsets: a training set (60% of the total data) and a testing set (40% of the total data). Table 4.3 presents the distribution summary and Variation Inflation Factor (VIF) of the variables in both datasets. The VIF is used to check for multicollinearity issues between variables. The maximum VIF is 3.89 in the training set and 3.14 in the testing set (far below the suggested threshold value of 10), indicating no significant multicollinearity issue within the datasets (Xu et al., 2017).

Table 4.3 - Statistic description of data

	Training dataset[1] (N=943 observations)					**Testing dataset[1] (N=624 observations)**				
	Min	**Max**	**Mean**	**SD**	**VIF**	**Min**	**Max**	**Mean**	**SD**	**VIF**
Sales rank (in the electronics category)	4	434,734	117,177	143,940	1.35	26	434,734	118,179	146,179	1.32
Price difference (in fraction)	0	1	0.26	0.24	1.03	0	0.95	0.26	0.24	1.04
Product promotion rate (in fraction)	0	0.55	0.01	0.05	1.02	0	0.5	0.01	0.04	1.03
Positivity of product description	0	13	4.69	3.8	1.10	0	12	4.53	3.76	1.17
Number of product pictures	1	7	3.49	2.29	1.05	1	7	3.56	2.34	1.02
Overall product rating	1	5	3.78	0.68	1.43	1	5	3.78	0.7	1.41
Number of service successes	0	158	45.71	56.20	3.89	0	158	46.54	56.29	3.14
Number of service failures	0	80	21.53	27.07	3.48	0	80	22	27.79	2.58
Number of questions answered	0	135	31.09	38.5	2.30	0	135	30	38.29	2.81
Total number of helpful votes	0	54	11.00	13.87	1.05	0	54	10.81	14.32	1.05
Average sentiment of customer reviews	-3	3	0.65	0.84	1.26	-3	3	0.65	0.86	1.28
Stock information	Yes (332); No (611)					Yes (45); No (579)				
Warranty information disclosure	Yes (80); No (863)					Yes (65); No (559)				
Brand equity	High (625) ; Low (318)					High (426) ; Low (198)				

[1] *Sales ranks were recorded on 30 May 2018, while all predictors were recorded in the period up to 30 April 2018 (i.e. one month before the May sales rank).*

4.7.1 Result of variable selection step

Regarding the result of the variable selection step, the minimal subset of eight strongly relevant predictors obtained by the RFE algorithm includes the number of answered questions, the number of service successes, the number of service failures, brand equity, the number of helpful votes, the overall product rating, the positivity of the product description, and the average sentiment of the customer reviews. The Boruta algorithm returns a subset of eleven all-relevant predictors, which extends the RFE subset with three weakly relevant predictors: the price difference, the number of product pictures and the stock information. Both the RFE and Boruta algorithms agree that warranty information disclosure and product promotion rates are not relevant predictors of remanufactured product demand.

4.7.1.1 Result of RFE algorithm

RFE employs a greedy optimisation algorithm to exhaustively search for the smallest subset of strongly relevant predictors with reasonable prediction accuracy. It iteratively fits a random forest to rank variable relevance based on permutation importance measures, and then successively drops the single least important variable at each iteration, until the smallest subset with high prediction accuracy is retained. The predictive performance (i.e. RMSE and R^2) of each subset, together with the corresponding standard deviation (SD), are summarised in Table 4.4 and plotted in Figure 4.2.

As seen in Table 4.4, the best predictive performance (RMSE = 0.0127 and R2 = 0.3267) is reached with twelve predictors. However, moving within one SD, the author can acquire the smallest subset available (i.e. comprising eight predictors), while retaining high prediction accuracy (RMSE = 0.0127 and R^2 = 0.3195). Such aggressive variable selection is based on the framework of gene selection by Díaz-Uriarte & Alvarez de Andrés (2006). The eight most relevant predictors for demand forecasting of remanufactured products selected by the RFE algorithm are highlighted **in bold** in Table 4.4.

Table 4.4 - RFE ranking of variable relevance based on predictive performance

Subset size	Variable name	RMSE	R^2	SD_{RMSE}	SD_{R2}
1	**Number of answered questions**	0.0150	0.1052	0.0009	0.0334
2	**Number of service successes**	0.0144	0.1437	0.0010	0.0381
3	**Number of service failures**	0.0139	0.1895	0.0010	0.0586
4	**Brand equity**	0.0134	0.2462	0.0010	0.0682
5	**Number of helpful votes**	0.0133	0.2552	0.0010	0.0689
6	**Overall product rating**	0.0131	0.2832	0.0010	0.0599

7	**Positivity of product description**	0.0129	0.2975	0.0010	0.0682
8	**Average sentiment of customer reviews**	0.0127	0.3195	0.0010	0.0716
9	Price difference	0.0128	0.3140	0.0010	0.0642
10	Number of product pictures	0.0127	0.3196	0.0010	0.0605
11	Stock information	0.0127	0.3264	0.0009	0.0582
12	Product promotion rate	0.0127	0.3267	0.0009	0.0566
13	Warranty information disclosure	0.0127	0.3257	0.0009	0.0572

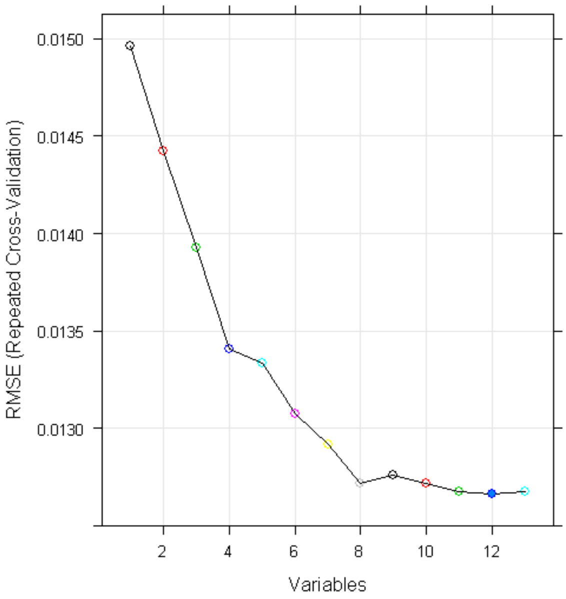


Figure 4.2 - RMSE reduction when adding and/or removing predictors using RF
(The x-axis is the number of variables used in the subset and variable names can be seen in Table 4.4)

4.7.1.2 Result of Boruta algorithm

The Boruta algorithm extends the dataset by duplicating all independent variables and then shuffles them to reduce correlation. The randomised variable of each original variable is called a shadow variable. The model determines whether a variable is important by comparing its Z-score (i.e. how many standard deviations a variable is away from its mean) with the maximum score of the best shadow attributes (shadowMax). The results of the Boruta algorithm are illustrated in Figure 4.3.

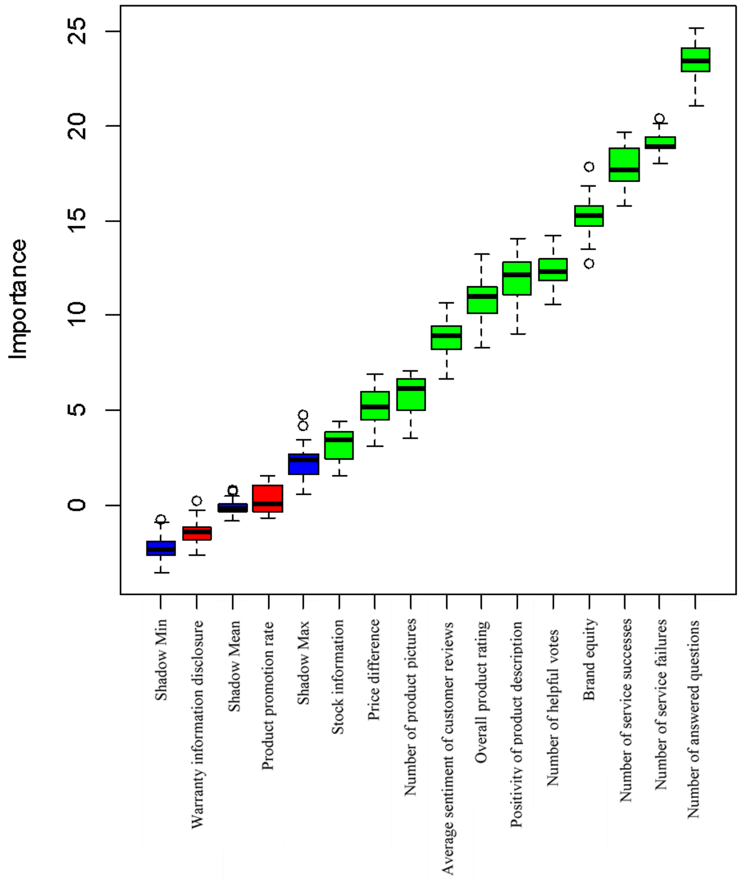


Figure 4.3 - Variable importance ranked by the reduction in Z-score (y-axis) of each independent variable (x-axis) in the Boruta algorithm

Variables with green boxplots are considered as relevant and selected. Variables with red boxplots are considered as irrelevant and removed. The minimum, mean and maximum score of the best shadow attributes, which serve as thresholds to make variable selection decisions, are represented by blue boxplots

Among the thirteen variables, warranty information disclosure and promotion rate are tagged as unimportant predictors (red boxplot) with Z-scores lower than shadowMax. Eleven variables with green boxplots are confirmed as important and selected as all-relevant predictors for RPDF.

4.7.2Model predictive performance evaluation

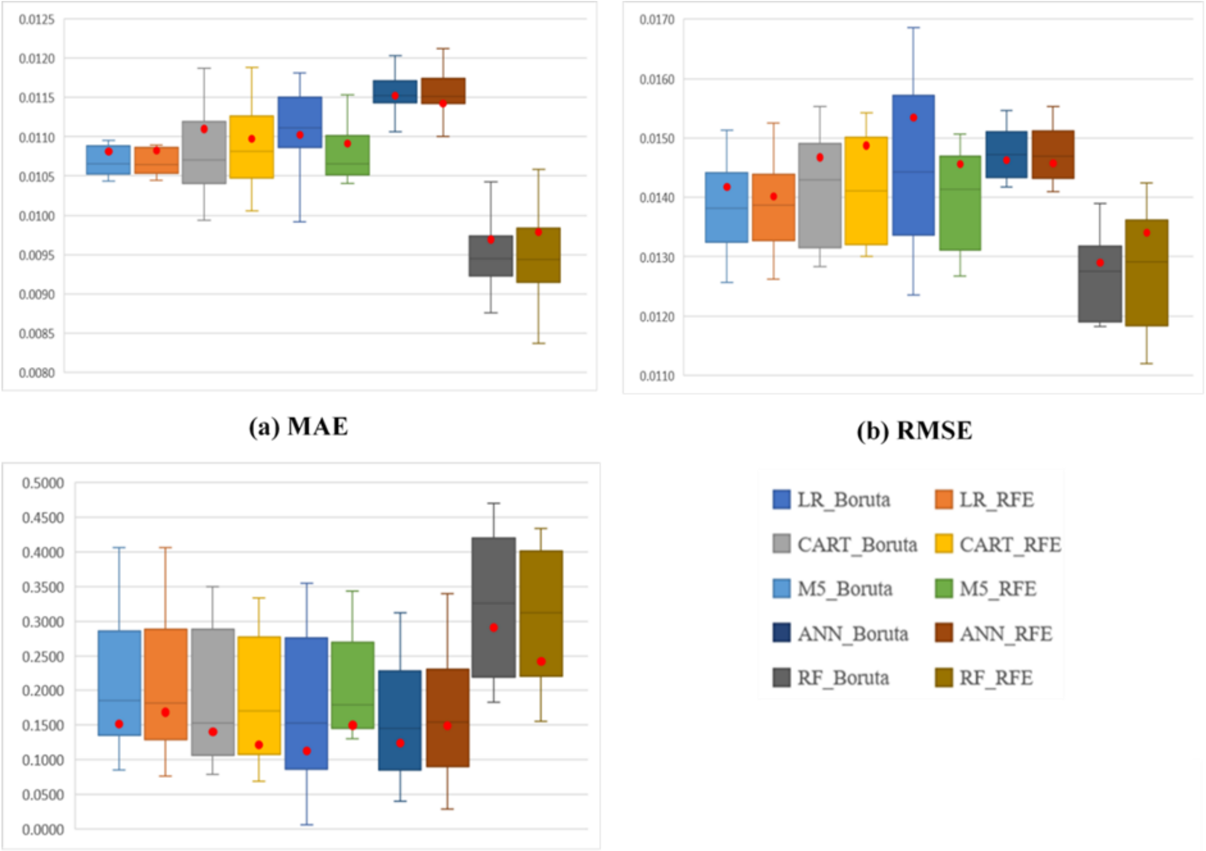
In this research, all the algorithms are fitted into the training dataset and then validated with the testing set. A ten-fold CV repeated three times is applied throughout the modelling process in order to remove the resampling bias.

To avoid bias issues, five algorithms are used to train the prediction model of remanufactured products: a linear model (i.e., linear regression, LR); two non-linear, single tree-based ML models (i.e., CART and M5); a non-linear, tree-based, advanced ML model (i.e., RF); and a non-linear, non-tree-based, advanced ML model (i.e., the artificial neural network, ANN). In LR, the input-output relationship is modelled using the linear predictor functions. ANN, inspired by the human brain, typically consists of input, hidden and output layers connected by processing units, so-called neurons. Neurons between layers are connected by the synaptic weights, and the ANN learning algorithm updates their weights to map the relationship between the predictors and the target variable. The sum of the weighted predictors is applied to the activation function in order to generate the prediction value (Witten et al., 2011). Due to the limited space, the description of the LR and ANN model can be referred to Witten et al. (2011). The model performance is assessed according to three measurements: MAE, RMSE and R^2. For a fair comparison, each model is run with its optimal hyperparameter settings and for both the RFE and the Boruta predictor sets, as shown in Table 4.5. Optimal parameter setting is also selected for both the Boruta and RFE input sets.

Table 4.5 - Model hyperparameter tuning based on ten-fold cross-validation

Predictive algorithm	No. of models per algorithm	Tuning parameter[2]	Candidate parameter setting[3]	Parameter selection [4] (RFE[5])	Parameter selection [4] (Boruta[6])
CART	10	Maximum depth of the tree	{1, 2, 3, 4, 5, 6, 7, 8 ,10, 11}	3	7
M5	8	Pruned	{Yes, No}	Yes	No
		Smoothed	{Yes, No}	No	Yes
		Rules	{Yes, No}	Yes	No
RF	10	Number of variables at each split	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11}	3	3
ANN	20	Number of hidden units	{1, 3, 5, 8, 10, 11, 13, 15, 17, 20}	10	10
		Weighting decay	{0.5, 0.1}	0.1	0.1
LM	1	N/A	N/A	N/A	N/A

[1] *A set of parameter values are created, and the algorithm is run for all possible value combinations between these tuning parameters.*
[2] *The tuning parameters of each predictive model can be specified using 'Caret' packages in R.*
[3] *The candidate parameter setting is randomly generated using the tuneLength() function.*
[4] *A parameter setting that can simultaneously optimise three performance measurements (i.e. RMSE, R^2 and MAE) is selected.*
[5] *Models are run with the eight most strongly relevant predictors obtained from the RFE algorithm.*
[6] *Models are run with the eleven all-relevant predictors obtained from the Boruta algorithm.*



(a) MAE

(b) RMSE

(c) R^2

Figure 4.4 – Prediction results in the training (boxplots) and testing (red points) stage with the ten-fold cross validation

The prediction results in the model training and testing phase are shown in Figure 4.4. In this figure, the boxplots represent the statistical distribution (min, mean and max) of each performance metric which results from training data with the ten-fold, three-time repeated CV, while the red point in each boxplot is the model performance based on the testing dataset in the model validation phase. In both the training and validation phase, it can be seen that RF outperforms the other predictive algorithms with the lowest values for errors and the highest R^2. It can also be seen in Figure 4.4 that all the values of MAE, RMSE and R^2 in the testing phase (the red points) fall within their corresponding distributions in the training phase (the boxplots). This indicates that there is no significant under-/overfitting problem. This validates the proposed approach to demand prediction.

Furthermore, the t-test is performed in order to compute the significance of the difference in performance measures (i.e. MAE, RMSE and R2) between the prediction algorithms used. The significance of the difference is measured with p-value. The difference is considered to be significant if p <0.01. Table 4.6 shows the table of the pairwise significance scores for MAE. The upper diagonal of the table provides the p-value for each pair of MAE scores. The lower diagonal of the table shows the estimated difference of each pair. Likewise, Tables 4.7 and 4.8 show the t-test results for RMSE and R2, respectively. As all three tables show, the p-values between RF_Boruta and other models (except RF_RFE), as highlighted in bold and underlined, are statistically significant. This reaffirms that RF outperforms the other models, as illustrated in Figure 4.4.

RF outperforms the LR model, indicating that remanufactured product demand is highly non-linear, a finding which requires the use of a non-parametric approach. Another interesting finding is that there is no significant improvement in model performance when using the Boruta predictor sets, compared to the RFE predictor sets. This implies that the predictive power of price difference, the number of product pictures and stock information on customer demand of remanufactured products are weak.

Table 4.6 - The t-test result for MAE

	(1) LR_ Boruta	**(2) LR_ RFE**	**(3) CART_ Boruta**	**(4) CART_ RFE**	**(5) M5_ Boruta**	**(6) M5_ RFE**	**(7) ANN_ Boruta**	**(8) ANN_ RFE**	**(9) RF_ Boruta**	**(10) RF_ RFE**
(1)	N/a	-3.14e-05	-9.64e-05	-1.19e-04	-5.07e-04	-5.28e-05	-9.85e-04	-1.01e-03	**1.16e-03**	1.17e-03
(2)	3.13e-01	N/a	1.00e+00	-8.71e-05	1.00e+00	-2.14e-05	7.61e-06	-9.80e-04	**1.59e-06**	1.20e-03
(3)	1.00e+00	6.50e-05	N/a	-2.21e-05	-4.11e-04	4.36e-05	-8.89e-04	-9.15e-04	**1.26e-03**	1.27e-03
(4)	1.00e+00	1.00e+00	1.00e+00	N/a	1.00e+00	6.57e-05	9.08e-04	7.90e-04	**1.78e-07**	1.29e-03
(5)	1.00e+00	4.76e-04	1.00e+00	3.89e-04	N/a	4.54e-04	-4.78e-04	-5.05e-04	**1.67e-03**	1.68e-03
(6)	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	N/a	2.80e-04	2.06e-04	**3.67e-06**	1.23e-03
(7)	5.51e-06	9.54e-04	1.26e-03	8.66e-04	1.00e+00	9.32e-04	N/a	-2.69e-05	**4.90e-09**	2.16e-03
(8)	6.07e-06	7.42e-06	1.11e-03	8.93e-04	1.00e+00	9.59e-04	1.00e+00	N/a	**6.94e-09**	2.18e-03
(9)	1.67e-06	-1.19e-03	8.42e-08	-1.28e-03	1.03e-04	-1.21e-03	-2.15e-03	-2.17e-03	N/a	1.23e-05
(10)	1.26e-05	9.72e-06	2.88e-07	9.98e-08	3.64e-04	8.14e-06	1.74e-08	1.74e-08	1.00e+00	N/a

Table 4.7 - The t-test result for RMSE

	(1) LR_ Boruta	(2) LR_ RFE	(3) CART_ Boruta	(4) CART_ RFE	(5) M5_ Boruta	(6) M5_ RFE	(7) ANN_ Boruta	(8) ANN_ RFE	(9) RF_ Boruta	(10) RF_ RFE
(1)	N/a	-3.59e-05	-2.40e-04	-1.66e-04	-5.75e-04	-1.04e-04	-9.64e-04	-9.87e-04	**1.14e-03**	1.13e-03
(2)	1.00e+00	N/a	1.00e+00	-1.30e-04	1.00e+00	-6.77e-05	1.19e-03	-9.51e-04	**5.20e-06**	1.17e-03
(3)	1.00e+00	2.04e-04	N/a	7.44e-05	-3.35e-04	1.37e-04	-7.23e-04	-7.47e-04	**1.38e-03**	1.37e-03
(4)	1.00e+00	1.00e+00	1.00e+00	N/a	1.00e+00	6.21e-05	3.60e-02	2.86e-02	**5.55e-07**	1.30e-03
(5)	1.00e+00	5.39e-04	1.00e+00	4.09e-04	N/a	4.71e-04	-3.89e-04	-4.12e-04	**1.72e-03**	1.70e-03
(6)	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	N/a	7.88e-03	4.97e-03	**2.23e-05**	1.23e-03
(7)	8.61e-04	9.28e-04	1.93e-01	7.98e-04	1.00e+00	8.60e-04	N/a	-2.36e-05	**4.19e-07**	2.09e-03
(8)	7.53e-04	9.54e-04	1.71e-01	8.21e-04	1.00e+00	8.84e-04	1.00e+00	N/a	**4.21e-07**	2.12e-03
(9)	9.34e-06	-1.18e-03	1.04e-05	-1.31e-03	3.15e-04	-1.25e-03	-2.11e-03	-2.13e-03	N/a	-1.23e-05
(10)	1.90e-04	8.46e-05	8.92e-05	3.99e-06	2.52e-03	2.54e-04	9.99e-07	8.01e-07	1.00e+00	N/a

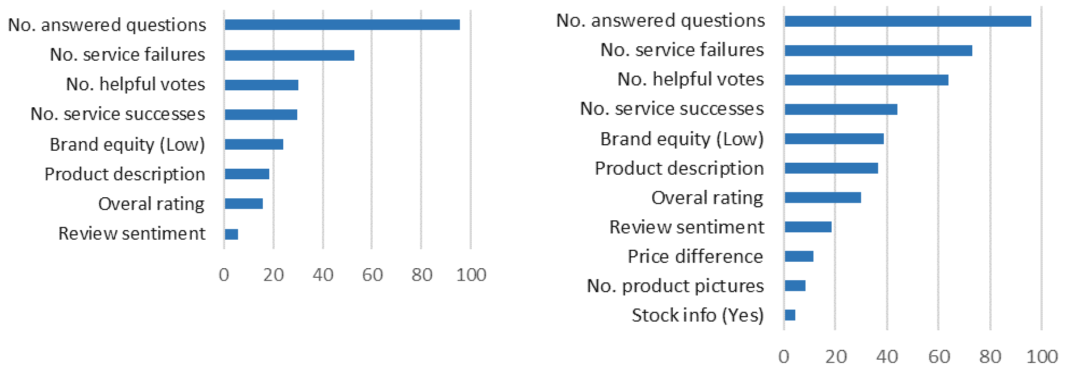
Table 4.8 - The t-test result for R^2

	(1) LR_ Boruta	(2) LR_ RFE	(3) CART_ Boruta	(4) CART_ RFE	(5) M5_ Boruta	(6) M5_ RFE	(7) ANN_ Boruta	(8) ANN_ RFE	(9) RF_ Boruta	(10) RF_ RFE
(1)	N/a	3.47e-03	1.53e-02	1.40e-02	2.60e-02	-3.24e-03	5.19e-02	4.91e-02	**-1.27e-01**	-1.21e-01
(2)	1.00e+00	N/a	1.00e+00	1.06e-02	1.00e+00	-6.71e-03	1.59e-02	4.56e-02	**2.63e-05**	-1.25e-01
(3)	1.00e+00	-1.19e-02	N/a	-1.30e-03	1.06e-02	-1.86e-02	3.66e-02	3.38e-02	**-1.43e-01**	-1.37e-01
(4)	1.00e+00	1.00e+00	1.00e+00	N/a	1.00e+00	-1.73e-02	1.00e+00	1.00e+00	**1.25e-06**	-1.35e-01
(5)	1.00e+00	-2.25e-02	1.00e+00	-1.19e-02	N/a	-2.92e-02	2.59e-02	2.31e-02	**-1.53e-01**	-1.47e-01
(6)	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	N/a	1.34e-01	7.81e-02	**1.75e-04**	-1.18e-01
(7)	2.25e-03	-4.84e-02	1.00e+00	-3.79e-02	1.00e+00	-5.51e-02	N/a	-2.80e-03	**3.60e-07**	-1.73e-01
(8)	4.22e-04	2.24e-03	1.00e+00	-3.51e-02	1.00e+00	-5.23e-02	1.00e+00	N/a	**3.54e-07**	-1.71e-01
(9)	3.10e-05	1.31e-01	5.86e-06	1.41e-01	7.93e-05	1.24e-01	1.79e-01	1.77e-01	N/a	5.97e-03
(10)	3.80e-04	2.51e-04	5.78e-05	5.33e-06	1.47e-03	9.64e-04	4.34e-06	2.33e-06	1.00e+00	N/a

4.7.3Model deployment – interpretation of results and managerial insights

As Kunc & O’Brien (2018) suggest, this research adopts a multi-methodological approach that combines different predictive and descriptive analytics tools to support the development of a marketing strategy for remanufactured products. After the prediction model is built in Section 4.5.2, the author further employs two descriptive analytics tools in order to gain data-driven marketing insights into remanufactured products. In particular, the VIR tool is used to identify the most influential market factors based on their predictive powers in the RF model. The PDP tool is then applied, in order to demonstrate the non-linear effect of these factors on customer purchasing behaviours. Such marketing insights can help marketers to understand the different ways in which customers of remanufactured products respond to marketing tools, compared to customers of new products, thereby developing an effective marketing strategy that can maximise the sales of remanufactured products.

4.7.3.1 Results and discussion of VIR analysis



(a)

(b)

Figure 4.5 - VIR result by using IFSA method with the RFE set (a) and the Boruta set (b)

In order to gain management insights from the RPDP model, the author first conducts the VIR analysis using the IFSA method (see section 4.6.2.1) for both the RFE (Figure 4.5a) and Boruta predictor set (Figure 4.5b). As can be seen in both figures, the number of answered questions, the number of service failures and the total number of helpful votes in customer reviews are ranked as the most important predictors of remanufactured product demand. The VIR also confirms that price difference, the number of product pictures and stock information are weakly important predictors, as explained in the previous section. The remaining predictors have moderately predictive powers on customer demand of remanufactured products. These include the number of service successes, brand equity (especially low brand equity), the overall product rating, the positivity of the product description, and the average sentiment of customer reviews. However, the importance ranking of variables is not enough on its own to provide managers with practicable insights. Therefore, a complementary analysis of PDP is necessary to examine the direction and magnitude of the impact of predictors on customer demand.

4.7.3.2 Results and discussion of PDP analysis

To extract the practicable insights from the RPDP model, the author analyses the PDP plots that illustrate the marginal effect of eight strongly and moderately important predictors of remanufactured product demand, as previously suggested by the VIR. It is notable that the PDP plots are drawn from partial dependence functions that are each populated with only a single predictor at any one time. As such, they show the nature of the relationship of a single predictor to customer demand after taking into account the average effects of all the other predictors in the model. While these plots do not fully represent the effect of each variable, they can serve as a useful basis for interpretation (Goldstein et al., 2015). In order to enhance the insights for management, the author also analyses the PDP contour plot representing the joint effect between two predictors on the response variable. As the RF model outperforms the other models in the prediction of remanufactured product demand, so the discussion in this section is based on the results of this model.

- Sales effect of the number of service failures and service successes

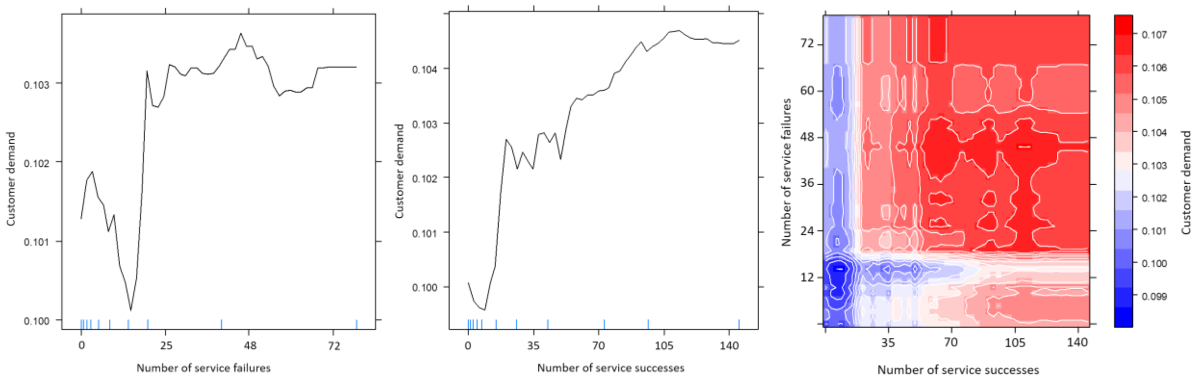


Figure 4.6 - The PDP of number of service failures and service successes

As with new products, the VIR analysis suggests that the number of service failures is more important than the number of service successes. The result is also supported by the PDP in Figure 4.6, which shows that the sales impact of service successes is more complex and non-linear than that of service failures. As a result, the used predictive algorithm (RF) captures less statistical (linear) correlation between it and customer demand. This can be explained by negativity bias: in other words, the psychological tendency for people to pay more attention, and give greater diagnostic weight to negative information than neutral and positive information.

Additionally, according to Park & Lee (2008), the number of service failures can play two roles in terms of social impact: an informant role, indicating the informativeness of customer feedback, and a recommender role, indicating product popularity. It is interesting at this point to examine how these two roles of the variable can influence customer purchasing behaviour related to remanufactured products. Interestingly, the PDP shows that when only a small amount of negative customer feedback is available (between 4 and 15 cases of negative feedback), the variable has a significant and negative association with customer demand. However, the effect becomes fairly positive and more non-linear when more than 15 customers give negative feedback.

This finding can be explained using the elaboration likelihood model (ELM), which is a theory of social psychology that describes the ways in which the information process of online customer reviews changes customer purchasing behaviour (Ho and Bodoff, 2014). According to the ELM, when only a limited amount of negative feedback is available, customers are more motivated to engage in a thoughtful and effortful information process, meaning that they read through the review content to look for quality cues of the product before making a purchasing decision. As customers examine the negative feedback carefully to gain better product knowledge, the informativeness effect of this variable outweighs its popularity effect. This means that an increasing number of service failures will significantly reduce customer demand. The author also observes an interesting finding which indicates a positive effect of negative customer feedback on sales, as in the case of remanufactured products with a very low number of service failures (fewer than 4 cases of negative feedback). As the ELM suggests, customers are very likely to carefully evaluate the product attributes from different reviews, leading to enhanced product knowledge and higher confidence, which may accordingly translate into higher sales. In contrast, when a large amount of negative feedback (more than 15 cases) is available, the ELM suggests that customers tend to adopt a low-depth information process which uses non-content, peripheral cues, such as a numerical index of product popularity, in order to make their buying decisions. Therefore, in such cases, when the number of service failures increases, indicating higher product popularity, the sales of remanufactured products also increase. However, the PDP suggests that this positive effect is much less significant and more non-linear, compared to the case outlined above of a low number of service failures.

Likewise, the marginal effect of the number of service successes is also in line with the ELM, and can therefore be explained in a similar way. On average, its effect on remanufactured products with low popularity (less than 15 cases of positive feedback) is significantly positive, whereas the effect on those with high popularity is less significant because of the highly complex and non-linear variances.

The joint effect of the number of service failures and the number of service successes on sales of remanufactured products also deserves attention. The contour plot suggests that there are strong interactions between these two variables. More particularly, remanufactured products with a high number of service failures (more than 15 cases of negative feedback) are associated with high customer demand only if there is also a high number of service successes (more than 15 cases of positive feedback), and vice versa. This finding highlights the recommender role of these two variables, showing that they are important indicators of product popularity. By increasing the amount of customer feedback, the seller can signal a strong awareness effect about the existence of the product, thereby placing it in the choice set of potential customers.

- Sales effect of overall ratings and customer review sentiments

Overall ratings and review sentiments are both representations of crowd intelligence. They are therefore perceived as trustworthy sources of information and are closely related to customer perception. The VIR result shows that the sales impact of overall ratings is more influential than that of review sentiments. This contradicts previous findings for new products which suggest that review sentiment has a more direct and substantial effect on demand than overall ratings (Hu et al., 2014). The result is supported by rationality boundary theory, which suggests that customers often seek to reduce their cognitive efforts by making product evaluations and purchasing decisions based on information that takes less effort to process and align, such as numerical ratings, rather than more effortful strategies such as reading textual customer reviews (Shah and Oppenheimer, 2008).

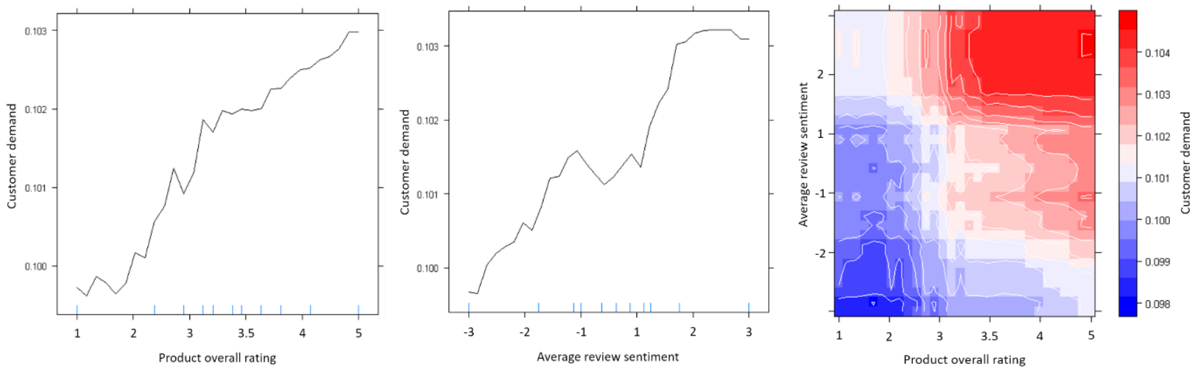


Figure 4.7 - The PDP of overall ratings and customer review sentiments

As Figure 4.7 demonstrates, the PDP shows that the overall product rating has a monotonic positive association with remanufactured product demand. This means that there tends to be higher demand for remanufactured products with higher overall ratings, which indicate positive seller reputation and goodwill accumulated over a long period. However, the magnitude of this impact varies greatly over the rating value range. This effect is moderately significant for remanufactured products with overall ratings above 3.5 stars; very significant for those rated between 1 and 3 stars; and mostly levelled off for those with less than 2 stars. Like the number of service failures, this finding can also be explained using negativity bias: i.e. unfavourable product ratings are likely to have a greater effect on purchase intention than favourable ones.

In contrast, the effect of review sentiments on remanufactured product demand is, on average, more non-linear in a heterogeneous way. In particular, the impact is relatively linear and significantly positive for remanufactured products with customer reviews which indicate negative polarity. For products with customer reviews indicating a positive polarity, the sales impact of the variable is significantly positive. However, such an effect is levelled off if the customer reviews indicate a strongly positive polarity. Finally, customer demand for remanufactured products with reviews indicating neutral polarity becomes very complex and non-linear. This is because prospective customers often need more time and cognitive effort to process and judge product reviews which have neutral sentiments (i.e. neither positive nor negative). Following rationality boundary theory, they therefore tend to be more risk-averse to remanufactured products which trigger great debate and require greater cognitive effort to evaluate.

The joint effect of overall ratings and review sentiments also provides some interesting insights for management. The contour plot shows that these two variables interact strongly in a heterogeneous way to affect customer demand for remanufactured products. It further suggests that remanufactured products with good reputations indicated by high overall ratings (3 to 5 stars) do not necessarily produce high sales, unless the average sentiment based on customer textual reviews also indicates a positive polarity. This is because only using numerical ratings cannot fully capture the information embedded in the product reviews. Indeed, by compressing complex, text-based customer reviews to a single number, the product is implicitly assumed to be one-dimensional, even though the economic theory for product differentiations posits that products are compounded with multiple attributes and each attribute has a different perceived importance to consumers, based on their individual preferences. Therefore, prospective customers do not simply base their decisions on the numerical overall ratings in order to reduce their cognitive effort, but also read textual reviews to gain more detail-rich information about the experiences, feelings and emotions of past customers.

- Sales effect of product description positivity

Of the three quality cues designed by online sellers (textual product descriptions, product pictures and warranty information disclosure), the VIR suggests that the textual product description is the most effective signalling tool with which to convey the quality and condition of remanufactured products. In contrast to new products, this study suggests that the product description is one of the more important quality cues for customers of remanufactured products. As a result, it is ranked as a moderately important predictor of demand in the VIR analysis. This may be because seller-generated product descriptions of new products are static and do not generally include intangible features, such as product quality, robustness, performance and so on. They may therefore be unimportant determinants of customers' purchasing decisions. However, such intangible product features are often contained in the sellers' descriptions of remanufactured products – with descriptions such as 'like-new', 'certified' and 'no scratches'

– in order to directly signal the condition of the product to customers. In this way, sellers can counteract the asymmetric information of the buyers and their perceived uncertainty, leading to higher sales, as suggested by market signalling theory.

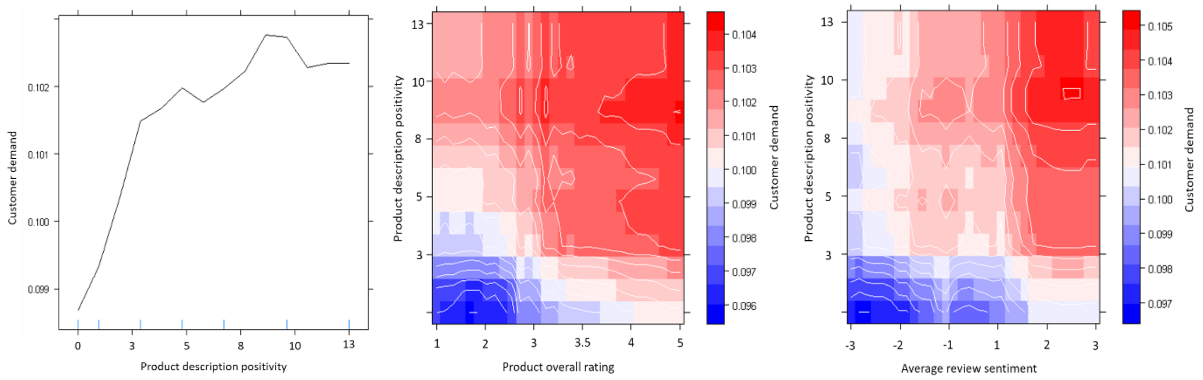


Figure 4.8 - The PDP of number of product description positivity

Regarding the marginal effect of this variable, the PDP in Figure 4.8 shows that, on average, the positivity of product descriptions has a positive association with sales. This effect is significant and linear for remanufactured products with descriptions indicating low positivity (i.e. those that contain fewer than three positive quality-related keywords). However, such an effect becomes moderate and relatively non-linear for RPs with strongly positive descriptions (containing between 4 and 9 such keywords). Nevertheless, as product descriptions are seller-generated rather than consumer-generated, the excessive use of this marketing tool may increase customer scepticism. As a consequence, the PDP shows that the variable has no effect or even a negative effect on sales when the product descriptions contain too many positive keywords (more than 9).

For practical uses, the author uses the PDP contour plots to examine how the positivity of product descriptions interacts with the other strongly and moderately important predictors of remanufactured product demand (excluding brand equity, as it is a categorical variable). The results suggest that the interactions between product description positivity and most of the consumer-generated variables are weak, and that increasing the positivity of product descriptions does not increase sales. See Figure 4.8 (right) for an example of its weak interaction with average review sentiment. However, this variable is found to have a strong interaction with overall product ratings, especially in the case of negative product ratings (lower than 3 stars). As Figure 4.8 (middle) suggests, when remanufactured products suffer low demand because of low product ratings which indicate a negative reputation, managers can raise sales to the average level by using product descriptions (ideally containing between 4 and 9 positive quality-related keywords) with strong implications about the high quality of the product.

- Sales effect of number of answered questions, number of helpful votes, and brand equity

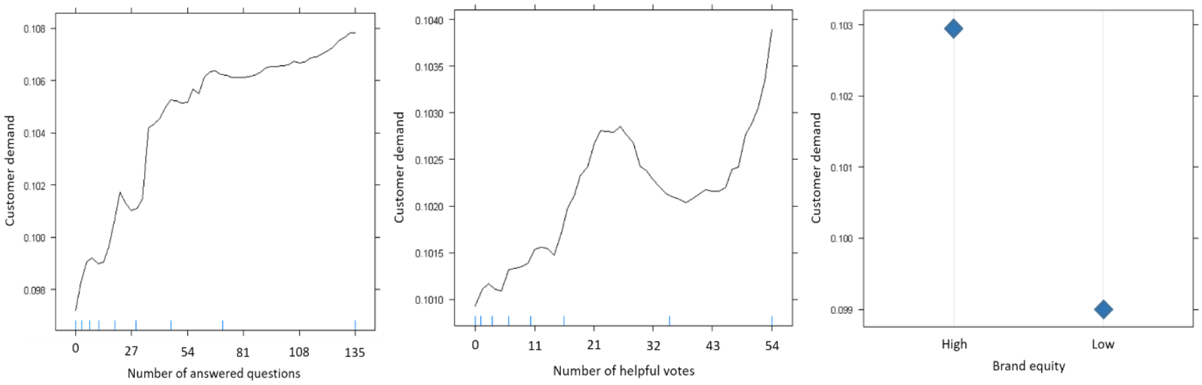


Figure 4.9 - The PDP of number of answered questions, number of helpful votes and band equity

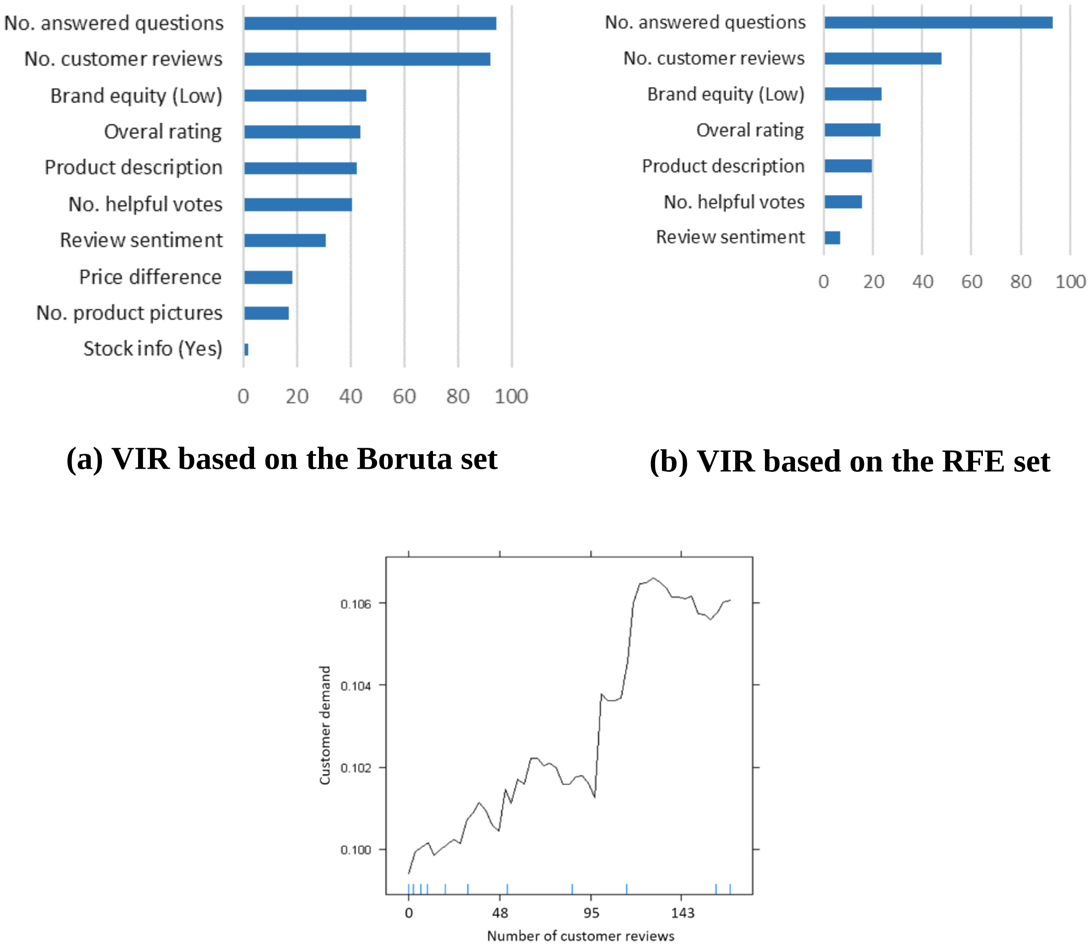
The VIR results suggest that the number of answered questions plays an important role in predicting customer demand for remanufactured products. More specifically, the PDP in Figure 4.9 shows that, on average, this variable has a monotonic positive influence on customer demand. Such an impact is significant when the number of questions ranges from 0 to 40. According to trust transference theory, when customers post more questions about a product, they stimulate social interactions and information exchange in the online community, which can help to ease the burden of information asymmetry, increase online trust and ultimately lead to higher demand. However, the positive effect of this variable becomes less significant when its value is high (more than forty questions answered). According to bounded rationality theory, customers' rationality is restrained by their cognitive limitations. They are therefore more likely to avoid information that requires a large amount of cognitive effort, such as when too many answered questions are available.

According to make signal theory, the number of helpful votes indicates the credibility of the review and this has a positive link with customers’ online trust and purchase intentions. Previous research has therefore found a significant and positive effect connecting this variable with sales of new products (De Maeyer, 2012). For remanufactured products, the results show that its predictive power of demand is relatively high and that the impact on sales is generally positive but in a non-linear fashion. More specifically, based on the PDP in Figure 4.9, the impact on sales is significantly positive when the perceived credibility of product reviews is either low (fewer than 21 helpful votes) or high (more than 37 helpful votes). However, when review credibility is at an average level (between 21 and 37 helpful votes), the impact on sales becomes negative. In signalling theory, over such a range of values, the variable sends out mixed signals (neither high nor low) about review credibility, which often requires prospective customers to make a greater cognitive effort in order to arrive at a judgement. However, since the cognitive resource of customers is limited, as rationality boundary theory suggests, they tend to become more risk-averse and less attracted to products which require a greater effort to evaluate. This accordingly has negative effects on sales. The PDP result also supports a positive link between brand equity and sales of remanufactured products, so that products with high/low brand equity are associated with high/low sales. In addition, the VIR suggests that brand equity is a moderately important predictor of remanufactured product demand, especially for those with low brand equity. Following market signalling theory, the brand equity of the OEM can be perceived as an initial cue signalling the reliability and quality of the product, justifying its perceived attractiveness and WTP.

4.8 Robustness check

4.8.1 Test 1: Using the number of customer reviews

The author replaces separate measures of product popularity and seller reputation (i.e. the number of service failures and the number of service successes) with a single measure: the total number of customer reviews. The author does not find any significant change in the prediction performance. As seen in Figure 4.10, the VIR shows that the new variable replaces the number of service failures as one of the three most important predictors. The importance ranking and PDP for most other variables remains the same. Unlike the number of service failures and the number of service successes, the number of customer reviews only indicates the popularity of the product (performing the recommender role), so its marginal effect is less complex than theirs. In line with signalling theory and the ELM, its sales impact is significant and positive for low popular remanufactured products, and becomes nonlinear and insignificant for high popular remanufactured products.



(a) VIR based on the Boruta set

(b) VIR based on the RFE set

(c) PDP based on RF_Boruta

Figure 4.10 – Result of robustness test 1

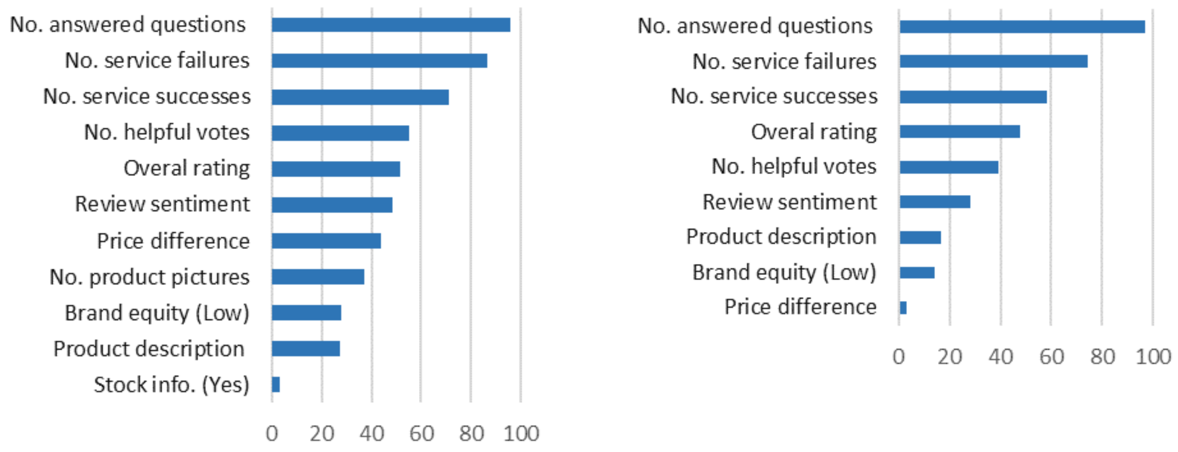
4.8.2 Test 2: Using the new dataset in different time periods

In the previous sections, the author uses the data of 1,567 remanufactured products listed on Amazon between April and May 2018. For a robustness check, the author uses the tracking crawler to collect new data for these products over the following six months. In particular, the sales rank was recorded up to 15 November 2018, while the lagged market factors were recorded up to 15 October 2018. There are 1,200 among 1,567 products available over this period. The author finds results that are consistent with the April-May data with regard to the predictive performance, VIR and PDP.

4.8.3 Test 3: Using different types of remanufactured products

To check whether the proposed approach for RPDP is valid across different product categories, the author collects a new April-May dataset that includes remanufactured products from the Home & Kitchen category in Amazon. This is the second largest category for remanufactured products after Electronics. It includes a wide range of small household appliances and kitchen equipment, such as food processors, coffee makers, mixers and vacuum cleaners. Previous CLSC literature distinguishes these remanufactured household products from the remanufactured technological products using the product repulsive level. Product repulsion refers to a customer's irrational and ingrained belief that pre-owned products are permanently tainted and therefore repulsive (Abbey, Meloy, et al., 2015). This means that technological products such as remanufactured laptops are "around-you" products which are associated with low repulsion. Household products such as remanufactured food processors are "on-you" products with medium to high repulsion. Personal care products such as remanufactured toothbrushes are "in-you" products which are associated with high repulsion. A more detailed classification of remanufactured products based on product repulsion can be found in Abbey, Meloy, et al. (2015).

The author finds 967 remanufactured household products over the period between April and May 2018. The RPDP approach described above is used to predict the May log sales rank based on the lagged market factors in April. Regarding predictive performance, RF still outperforms the other model with the highest prediction. The results show that eleven of the thirteen independent variables (excluding brand equity and product description positivity) have a consistent impact, as in the case of technological products. This means that the product category and product repulsion on customer purchasing behaviour of remanufactured products did not have pronounced effects.



(a) VIR based on the Boruta set **(b) VIR based on the RFE set**

Figure 4.11 – Result of robustness test 2

However, there are still some interesting results that deserve attention. Based on the VIR in Figure 4.11, household remanufactured products contrast with technological remanufactured products, as brand equity and product description positivity have the lowest predictive powers, indicating that customers do not perceive them as cues signalling quality. This is because when customers hold a high repulsive perception, as in the case of household products, the perceived risks and uncertainty about the quality of the remanufacturing process become more severe. This means that they are more reliant on consumer-generated variables reflecting crowd intelligence than on seller-generated variables as trustworthy sources of information. This is why all consumer-generated variables for remanufactured household products have moderate to significant predictive power of customer demand, as shown in the VIR.

A summary of the main findings based on the VIR and PDP analysis is presented in Table 4.9.

Table 4.9 - Summary of main findings in this research

Predictor	Signal Implication	Predictive power of RP demand	Management implications of the findings on remanufactured product (RP) demand	Previous findings on new products
Number of questions answered	Social interaction	Strong	(1) The effect is monotonic positive. (2) Such an effect is more significant when the number of answered questions ranges from 0 to 40, and less significant when there are more than 40 questions answered.	Significantly positive
Number of Service failures	Seller reputation, Product popularity	Strong	(1) For low popular products, the effect is significantly negative when there is only a small amount of negative feedback available (between 4 and 15), and becomes significantly positive when there are fewer than 4 cases of negative feedback. (2) For high popular products, the effect becomes somewhat positive and more non-linear when there is a high number of service failures (more than 15 negative feedback). (3) The impact of service failures is more important than service successes.	Significantly negative
Number of helpful votes	Review credibility	Strong	On average, the effect is significantly positive, except for products which indicate neither very low nor very high review credibility (i.e. 21 to 37 helpful votes) for which the effect becomes negative.	Significantly positive
Number of Service successes	Seller reputation, Product popularity	Moderate	(1) For low popular products, the effect, on average, is significantly positive when there are less than 15 cases of positive feedback. (3) For high popular products, the effect is highly complex and very non-linear, and hence statistically unimportant. (3) Products with the highest customer demand are those which have both a high number of negative feedback (more than 11 cases of feedback) and a high number of positive feedback (more than 15 cases of feedback), which indicate the maximum product popularity and customer awareness.	Moderately positive
Brand equity	Quality cue	Moderate (technological RP)	For technological products featuring a low customer repulsion level, the brand equity of the original manufacturer has a significant and positive effect on RP demand. This means that RP demand decreases when moving from high to low brand equity. The effect of low brand equity is also more pronounced than that of high brand equity.	Moderately positive
		Limited (Household RP)	For household products featuring medium to high repulsion level, the brand equity of the original manufacturer does not have significant effect on RP demand. Customers are more reliant on information generated by peer customers, such as customer reviews and ratings to make a decision.	
Overall product rating	Seller reputation	Moderate	(1) On average, the overall rating has a monotonic positive effect on sales, but the impact magnitude varies greatly. (2) The positive effect is moderately significant for products with positive ratings (more than 3.5 stars), very significant for those with neutral and negative ratings (between 2 and 3 stars), and mostly levelled off for those with very negative ratings (fewer than 2 stars). (3) However, its positive effect is largely dependent on the textual sentiments of customer reviews. For example, products with good reputation indicated by high overall ratings (3 to 5 stars) does not necessarily lead to high sales, unless the average sentiment based on customer textual reviews also indicates a positive polarity.	Moderately positive
Customer review sentiment	Seller reputation	Moderate	(1) This variable is less important than the overall rating because, on average, the effect of this variable is more non-linear than that of the overall rating in a heterogeneous way. (3) The effect is significantly positive for products with customer reviews which indicate negative polarity. (2) The effect is significantly positive for products with customer reviews which indicate positive polarity, but such an effect is levelled off for products with strongly positive polarity. (3) The effect is very complex and non-linear for products with customer reviews which indicate neutral polarity (neither negative nor positive polarity).	Significantly positive
Product description	Textual quality cue	Moderate (technological RP)	(1) The effect is significant and linear when the product descriptions indicate low positivity (containing fewer than three positive quality-related keywords); and becomes moderate and relatively non-linear with strongly positive product descriptions (containing between 4 and 9 such keywords. The effect is levelled off or becomes negative if the product descriptions contain too many positive keywords (more than 9). (2) The variable moderates the sales effect of the overall product rating. Therefore, when products have low demand because of low product ratings (fewer than 3-stars), managers can raise sales to the average level by using product descriptions (which ideally contain between 4 and 9 positive quality-related keywords).	Insignificant
		Limited (Household RP)	For remanufactured versions of household products, customers do not perceive the product descriptions provided by the sellers as reliable quality cues.	
Product pictures	Visual quality cue	Limited	Customers do not see the product pictures as important cues signalling the quality and condition of the RP.	Positive
Warranty information disclosure	Quality cue	None	Since few Amazon sellers voluntarily provide the full terms and conditions of warranties for the RPs, the effect of this variable on RP demand is unconfirmed in this study, but is worth investigating in future studies.	Positive

Price difference	Price discount	Limited	Using price incentives is not an effective way of increasing sales. Sellers should instead focus on improving their reputation and goodwill by obtaining a higher number of positive ratings from past customers and by becoming more actively involved in question-and-answer activities.	Positive
Promotion rate	Price discount	None		
Stock information	Limited availability	Limited	The psychological effect of the scarcity principle does not apply to RP customers; therefore, the selling strategy of using limited stock availability does not help boost sales.	Positive

4.9 Summary

This chapter develops a comprehensible data-mining approach in order to predict remanufactured product demand and to develop a marketing strategy. Based on the real dataset on www.amazon.com, the results show that the ensemble regression tree model, RF, can provide the most accurate and robust prediction result. A VIR analysis is then applied to determine the most influential market factors based on this prediction model. The effect of these factors on RP demand are examined using PDP analysis. As a result, a number of data-driven marketing insights are revealed, providing guidelines to help managers design an effective marketing strategy specific to remanufactured products, as summarised in Table 4.9.

The author believes that this research can be grouped with other pioneering studies which provide a structured framework showing how business analytics can be applied to support data-driven demand forecasting and marketing strategy development in the remanufacturing/CLSC literature. Future research should focus on three aspects. (1) The findings of this research rely on data from only one online marketplace, www.amazon.com. In practice, however, customer behaviour could change when comparing product deals from multiple online sales channels. Therefore, it may be beneficial to aggregate data from various sources. (2) ML models are often seen as black boxes, which limits their applications in industry. Future research should make a greater effort to develop new methodologies that can effectively explain the results of these apparently incomprehensible models. (3) The application BDA in reverse logistics and CLSCs is yet to be fully appreciated, and there is currently little research into these areas. This research is expected to serve as an example and stimulate further investigation.

Chapter 5 Data-driven prescriptive optimisation of promotional pricing

Addressing the questions of “what will happen” and “why will it happen” in the future, predictive analytics has well-proven to contribute considerably to the business values. However, it is still not possible to obtain the maximum potential of predictive analytics since there is a time interval between the event prediction and proactive decision, leading to an inevitable business value loss (Lepenioti et al., 2020). Therefore, to maximise the business value, practitioners often perform predictive analytics in conjunction with prescriptive analytics to optimise the proactive decision making ahead of time. Saying that, this chapter is considered as the extension of the predictive analytics model in Chapter 4 to generate prescriptive analytics that can fully exploit the data-driven predictions and reach the utmost business value. The extended approach is validated using a real case study of promotional price optimisation across the retail chain.

The chapter is structured as follows: the research background, aim and objectives are introduced in Section 5.1. The methodology is described in Section 5.2. The experiment setting to validate the model as well as its result is discussed in Section 5.3. The conclusion of this chapter is in Section 5.4.

5.1 Introduction

Recent progressions in ML have made a huge improvement on business efficiency in almost every industry. As found in the literature view chapter (Section 2.5), prescriptive analytics and predictive analytics have been the largest and fastest growing research areas in the BDA-driven SCM. Current advances in predictive analytics in terms of both algorithms and technologies have made it fairly straightforward to generate a large amount of accurate and timely predictions purely based on data. However, the key question is in prescriptive analytics which focuses on how to leverage those massive predictions to maximise the effectiveness and efficiency of the complex decision making process. The development of prescriptive analytics often raises a technical issue regarding the integration of ML with relevant theories and algorithms about mathematical optimisation and numerical simulation (Lepenioti et al., 2020).

There are two important outcomes that are usually generated from predictive analytics. The predictive formulas can (1) estimate predicted values for a specific key performance indicator (KPI), and (2) reveal the hidden relationship between dependent and independent variables. The data-driven predictive approach proposed in Chapter 4 above is a good demonstration of how to obtain both outcomes in the sales forecasting problem. However, as compared to descriptive analytics and predictive analytics, prescriptive analytics is particularly a critical step to help companies reach the utmost value out of the BDA adoption. One of the most straightforward and commonly used approach to obtain the prescriptive analytics level is to integrate the predictive analytics with the mathematical optimization by treating the forecasted value as inputs in the optimisation model (Lepenioti et al., 2020). There are a number of existing studies in various sectors, which have adopted this approach, for example, airline operations (Achenbach and Spinler, 2018), new product development (Dey et al., 2017), and patient scheduling (Srinivas and Ravindran, 2018). Hence, the aim of this chapter is to maximise business values of BDA by integrating the sales forecasting approach with the promotional price optimization model to achieve the prescriptive promotional price optimisation.

Consumer price promotion, which refers to temporary price discounts offered to customers, is an important sales strategy for retailers to increase the product demand from price sensitive customers (Bogomolova et al., 2017). Empirical research has shown that price promotion can have positive impacts not only on companies’ revenues and profits but also on their intangible assets such as brand loyalty and brand equity (Kuntner and Teichert, 2016). However, price promotion is also a complex and challenging decision making process, which would lead to adverse effects if it is poorly planned. This is because (1) the price promotion of a product affects not only the demand for the focal product but also the demand for the other products in the same store, meaning that the promotion can help boost sales of the focal product but probably at the cost of reducing profits arising from other products; (2) since manufacturers often provide promotion deals on certain products at certain time of the year, identifying the timing of the promotion is also a critical decision; (3) due to resource scarcity, price promotion is often constrained by a set of business rules specified by the company and/or products such as marketing budgets and number of promotions allowed for each product family; and (4) the problem becomes difficult than ever before as the retailer store has now provided a wide range of product types leading a large number of promotional pricing decisions that have to be made in the daily/weekly operational basis (Ma and Fildes, 2017). Given these complexities of the price promotional problem, a data-driven decision making approach is essential for the practical use.

Promotional price optimisation has been actively studied in marketing/OR literature (Caro and Gallien, 2012; Ferreira et al., 2016; Harsha et al., 2019; Ito and Fujimaki, 2017; Kunz and Crone, 2014). However, the practical use of these existing studies is rather limited due to the capability of the demand prediction model, i.e., most of the demand models are still based on the parametric approach with linearity assumptions (eg., Caro & Gallien, 2012; Harsha et al., 2019). Ferreira et al. (2016) is one of very few studies that adopt the nonparametric, ML-based approach to the prescriptive price optimisation; however, the demand prediction is based only on decision tree algorithm of which the prediction performance is not robust to the change of the market input, as explained in Section 4.5.2, Chapter 4 above.

Therefore, this chapter aims to develop a prescriptive, data-driven model of promotional price optimisation, which can derive the optimal pricing strategy to maximise the future sales on the basis of the more robust demand prediction approach that compares the performance of different ML models as proposed in the previous chapter. In particular, the proposed approach has two stages. First, the author generates accurate sales forecasts at product-store level, which unveils complex relationships between product sales and prices. Such information is then used as inputs in the second stage where the author develops the mixed-integer linear programming (MILP) model that optimises the promotion planning. In the optimisation model, the author aims to determine optimal promotional price of each product in order to maximise the retailer's total sales of the hole categories under while taking into consideration a set of constraints modelling important business rules.

To validate the approach, the author uses the real-world historical sales data collected by IBM Watson analytics, which includes 81,380 sales transactions of 21 product types sold on different sales channels in 19 countries between 2012-2014. The result shows that the proposed prescriptive approach fits the data well in the way that accurately forecasts the sales and optimise the promotional pricing planning of multiple products with the increase of the cumulative sales by 26% and revenue by 35%, as compared to the original pricing plan.

5.2 Methodology

This study aims to address the promotional pricing optimisation problem using the data-driven, prescriptive approach. In particular, the challenge is to help business planners design the upcoming promotion campaign (eg. during the Christmas season) by determining which products among a variety of product categories should be put on sale at what discounted prices, and the ultimate goal is to maximise the total sale of the retailer while obeying various business constraints.

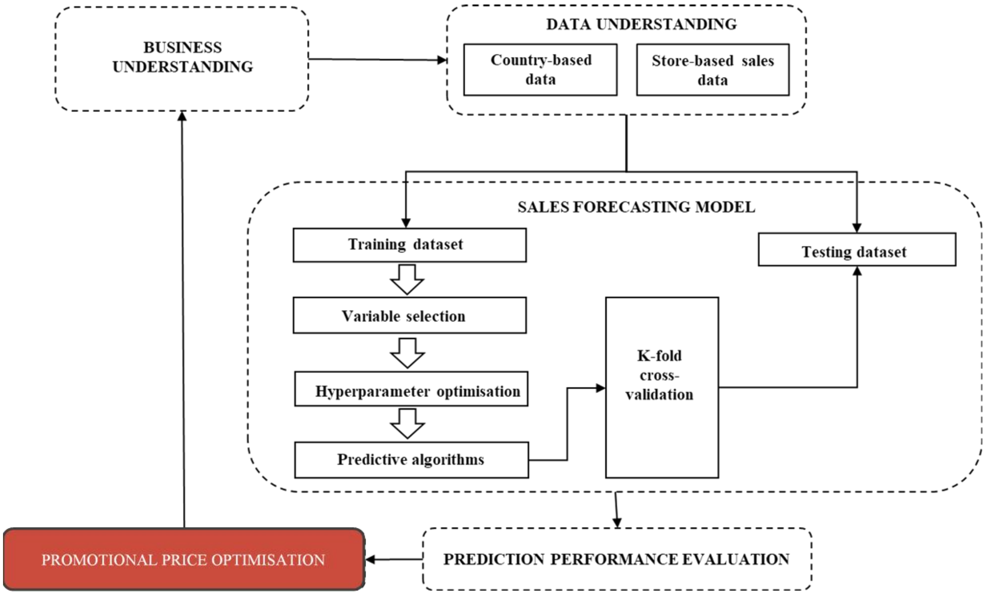


Figure 5.1 – Methodology framework

(Source: The author)

There are three types of variables in the prescriptive price optimisation problem, namely decision, target and external. Decision variables are ones that are optimised, i.e., promotional prices of products. Target variables are those we predict, i.e., sales quantities. External variables refer to other information that we can include in the model, for example, weather data, macroeconomic factors of countries, and product information, etc. The proposed prescriptive price optimisation is conducted in two stages. In the first stage, we need to predict the impact of different pricing levels on the expected sales for each type of products. To do so, different machine learning techniques are used to generate the prediction of the target variables (i.e., product sales) by employing the decision variable (i.e., product price) and external variables (i.e., weather data, macroeconomic factors of countries, and product information) as input features. By using ML-based predictive analytical approach, this stage can unveil the non-linear, complex relationships between sales and prices of multiple products, taking into account the price elastics of the product demand, product cannibalisation, as well as the effect of external information. In the second stage, the price-sales relationship is transformed into a mixed-integer linear programme (MILP) optimisation problem. Essential business rules are represented as linear constraints predefined by the user of the system. By solving the optimisation problem, the optimal values of the decision variables (product prices) are determined.

The overview of the proposed prescriptive price optimisation can be seen in Figure 5.1. For the sales forecasting at the first stage, the ML-based prediction approach proposed in Chapter 4 is used. At the second stage, the promotional price optimisation model is described as follows.

In this study, the price optimisation is formulated as a MILP model. The model is developed particularly for the class of non-perishable products with the assumption that there is a well-established inventory replenishment policy in place and that the stockout effect is negligible. This assumption is quite reasonable for non-perishable products (Harsha et al., 2019). Mathematically, it enables one to investigate the single period pricing problem in the promotion planning without the interference of inventory effects. For simplicity, the author also assumes that the substitute and complementary effects between multiple products in the retail chain are absent.

Based on the above assumptions, the study focuses on the single period promotional pricing optimisation problem for multiple products. The objective is to maximise total sales across the retail chain, whereas the decision variables are for each product line, what types of products are promoted and at what price. The mathematical formulation of the optimisation problem is described as follows.

Suppose we have I products indexed by $i \in \{1, \dots, I\}$ and they are from J product lines indexed by $j \in \{1, \dots, J\}$. Let P_i be the discrete set of admissible prices for product i constructed based on the historical data, and $G_i = |P_i|$ be the number of possible prices of product i. Then, we have $p_{i,g}$ representing the g-th possible price in set P_i, where $g \in \{1, \dots, G_i\}$. The prices in the P_i set are sorted in the descending order where the first price (i.e. $p_{i,1}$) is regarded as the normal price and the others (i.e. $p_{i,2}, \dots, p_{i,G_i}$) are regarded as the discounted prices. The decision variable of the optimisation model is defined by a binary variable $x_{i,g}$ such that $x_{i,g}$= 1 if product i is assigned to price $p_{i,g}$, and $x_{i,g}$= 0 otherwise. Also, let $\hat{d}_{i,g}$ be the forecasted sales of product i at price $p_{i,g}$, which is the target variable in the sales forecasting model in the first stage. Then, the objective function of the price optimization in the second stage can be expressed as follows:

$$Max_z \quad sales = \sum_{i \in I} \sum_{g \in G_i} \hat{d}_{i,g} x_{i,g} \tag{5.1}$$

where $sales$ is the objective variable, which is the total sales of all products during the promotional campaign.

When optimising the promotional pricing, there are various important practical business rules that planners should take into account. First, we need to ensure that each product can only be assigned to exactly one price among the price candidate set. Thus, we have the constraint:

$$\sum_{g \in G_i} x_{i,g} = 1 \quad \forall\, i \in I \tag{5.2}$$

Second, as the marketing budget allocated for the promotional campaign is often limited, the number of products concurrently under the price offers should be constrained. Too many or too few products with price reductions at the same time may counteract the promotional effect or even damage the retailer's brand image. Here, this rule can be expressed in terms of the following constraint:

$$\sum_{i \in I} \sum_{g \in G_i} x_{i,g} \leq max_j \;\; \forall j \in J \tag{5.3}$$

Where max_j is the predefined maximum number of products with discounted prices in each product line j.

The final constraint refers to the maximum revenue that the company is willing to give up in cases selling at lower prices for higher sales quantity might lead to the lower revenue. With proportional pricing, the company sometimes aims to offer the lower price with the expectation that it would lead to higher sales quantity and increased customer reach, though the revenue in the short-term may be reduced. However, excessive price reduction would damage the company profitability, hence decision makers are usually required to put a cap on how much revenue the company is willing to give up for the campaign. This business requirement can be reflected in the constraint below:

$$\sum_{i \in I} \sum_{g \in G_i} (\hat{d}_{i,g} - \hat{d}_{i,1})^{+} p_{i,1} x_{i,g} \leq cap \tag{5.4}$$

Where cap is the predefined maximum revenue that the company is willing to give up during the promotional campaign. Note that we use the positive part (+) in the constraint (5.4) because it is more likely that the expected sales of products with the discounted price $p_{i,g}$ are higher than those with the normal price $p_{i,1}$.

As the problem size is not too large, the proposed MILP problem from Eq (5.1) to Eq (5.4) can be solved using the off-the-shelf solver on Python to obtain the exact result.

5.3 Experimental result

5.3.1 Data understanding

The proposed prescriptive price optimisation is validated using the real-world retail dataset provided by IBM Watson Analytics (available on https://www.kaggle.com/samipjshah/wa-sales-products-201214). It includes 81,380 quarterly sales records of 21 product types and 5 product lines. The data is collected from a sample of retail stores in 19 countries over the period from Quarter 1, 2012 to Quarter 2, 2014, including information on sales, costs, profits, order methods, and retailer types.

Empirical research has well-recognised impacts of exogenous external variables such as weather and macroeconomic factors of store locations on sales (Sagaert et al., 2018). Therefore, the author enriches the sales data with additional location-specific features to integrate the context-awareness element into the sales forecasting model. In a real-world business application of ML, this step may be instructed by domain experts to identify what external features are relevant to the target variable. Here, the author employs the external variables that are commonly used for sales forecasting, namely weather data and retailer country-based macroeconomic data (Sagaert et al., 2018). For the weather data, the author collects the quarterly average temperature and humidity data between 2012 and 2014 of each capital city to represent the country weather data, using https://www.timeanddate.com/weather/. For the macroeconomic data on the country level, the author chooses three macroeconomic factors that can make countries similar or dissimilar to each other with regards to retail sales. Those features are Gross Domestic Product (GDP) per capita, Consumer Price Index (CPI) and unemployment rate. They are quarterly data between 2012 and 2014, collected from https://data.oecd.org. To ensure the model accuracy, the author applies the variable selection step by which all variables that are irrelevant to the target variable (i.e., sales) will be filtered out before the model training stage. Finally, the author joins the weather data and the macroeconomic data with the sales data by taking the country code as the joining key.

5.3.2 Data preparation

There are some preprocessing tasks that need to be done with the joint dataset before training the prediction model. The first step is data extraction in which new predictors can be extracted from the current data. When dealing with time series analysis, it often involves calculating the lagged value of the target variable to measure the auto-correlation effect, i.e., how past values of a variable influence its future values, hence revealing predictive values. Here, the author creates a lagged sales variable from the previous quarter as a predictor of the current sales. To take into account the seasonal effect, the author also employs a new categorical variable called "Quarter" which includes four quarters of the year. The second step in the data preparation is data transformation in which the continuous variables are normalised to the range between 0 and 1 to address the scale effect between them, while the categorical variables are transformed into the dummy variables to be treated as numerical variables. The last step of data preparation is to split the dataset into two subsets: training data and testing data. The joint dataset includes the quarterly sales, product-specific, and external features for 11 quarters during the 2012-2014 period. Hence, the first 8 quarters (Quarter 1, 2012 to Quarter 4, 2013) are used to train the prediction models, and the last 3 quarters (Quarter 1, 2014 to Quarter 3, 2014) are used to validate the prediction.

Table 5.1 presents the distribution summary and Variation Inflation Factor (VIF) of the variables in both the training and testing dataset. The VIF is used to check for multicollinearity issues between variables. The maximum VIF is 1.96 in the training set and 2.94 in the testing set (far below the suggested threshold value of 10), indicating no significant multicollinearity issue within the used datasets (Xu et al., 2017). Table 5.2 and Figure 5.2 summarise the distribution of sales transaction records by product type and retailer country, respectively.

Table 5.1 - Statistic description of data

	Training dataset (N= 54,519 observations)					**Testing dataset (N=26,861 observations)**				
	Min	**Max**	**Mean**	**SD**	**VIF**	**Min**	**Max**	**Mean**	**SD**	**VIF**
Quarter sales quantity *(Target variable)*	1	42,431	701	1,360	1.42	1	67,875	871	1,779	1.30
Previous quarter sales	3	13,577	505	505	1.46	3	13,577	537	563	1.38

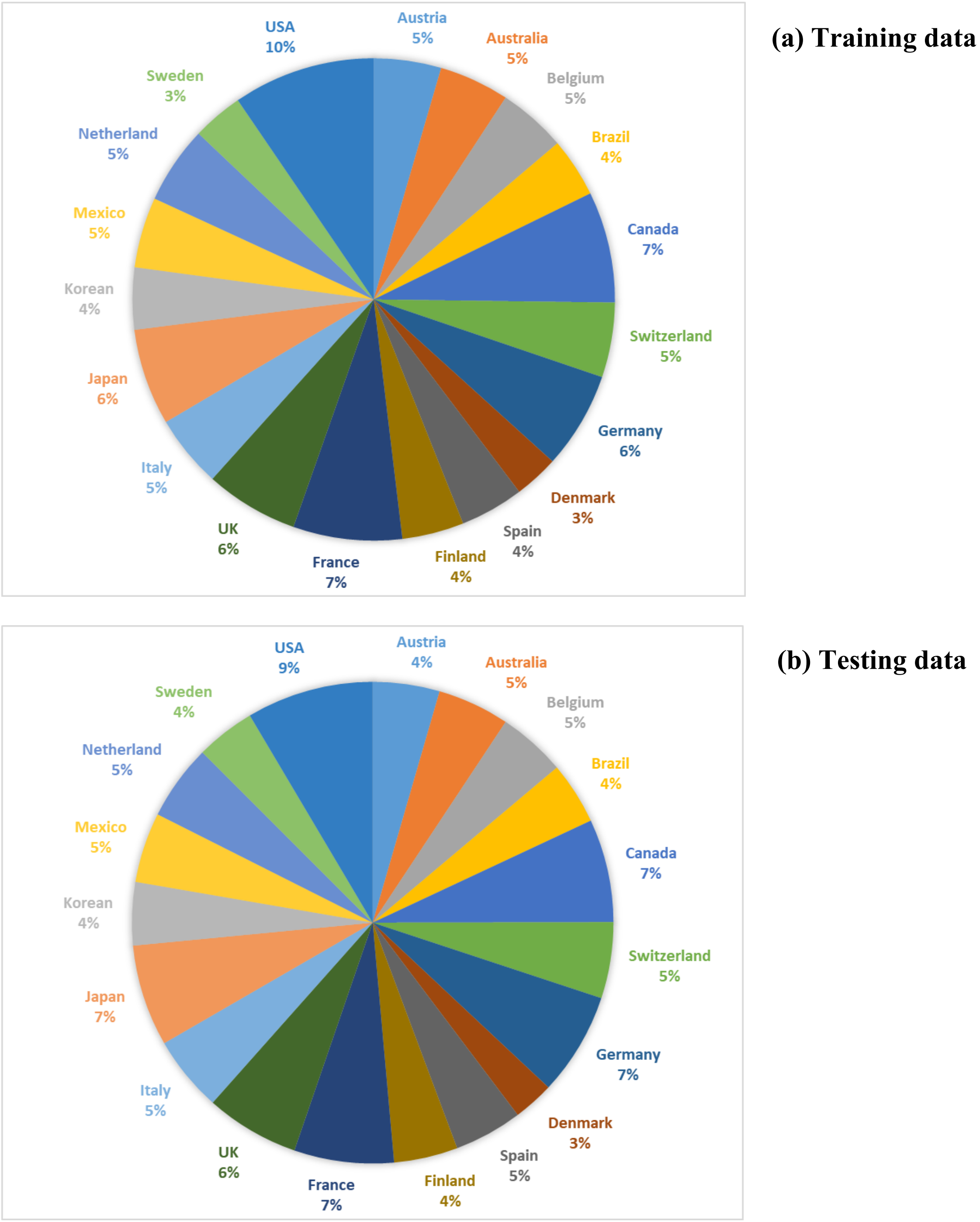
Selling price	1	1,305	137	209	1.07	1	1,305	136	208	1.08
GDP per capita[1]	15,120	60,566	40,791	9,875	1.96	15,374	62,266	42,242	10,434	2.27
CPI rate[1]	-1	7	2	1	1.66	0	7	2	1	2.94
Unemployment rate[1]	3.07	26.32	7.54	4.26	1.04	3.10	25.80	7.51	4.42	1.78
Average temperature[2]	-5	27	12	8	1.24	-9	26	12	7	1.33
Average humidity[2]	38	89	70	10	1.45	36	90	71	11	1.35
Quarter	**Q1** (15,742); **Q2** (15,660); **Q3** (15,623); **Q4** (7,494)					**Q1** (7,728); **Q2** (7,508); **Q3** (4,396); **Q4** (7,229)				
Order method type	**E-mail** (2,406); **Fax** (1,163); **Mail** (964); **Sales visit** (5,099); **Special** (368); **Telephone** (4,631); **Web** (39,888)					**E-mail** (686); **Fax** (296); **Mail** (31); **Sales visit** (1,844); **Telephone** (885); **Web** (23,119)				
Retailer type	**Department Store** (12,247); **Direct Marketing** (1,798); **Equipment Rental Store** (1,003); **Eyewear Store** (3,474); **Golf Shop** (4,970); **Outdoors Shop** (15,046); **Sports Store** (13,072); **Warehouse Store** (2,909)					**Department Store** (5,006); **Direct Marketing** (1,034); **Equipment Rental Store** (523); **Eyewear Store** (1,956); **Golf Shop** (2,712); **Outdoors Shop** (7,671); **Sports Store** (6,391); **Warehouse Store** (1,568)				

[1] *Quarterly average values of macroeconomic factors in the retailer countries*

[2] *Quarterly average values of weather data in the retailer countries*

Table 5.2 - Summary of product lines and product types by transaction counts

Product line	**Product type**	**Number of transactions**	
		Training data	**Testing data**
Camping Equipment	Cooking Gear	4,029	1,851
	Lanterns	4,350	2,102
	Packs	2,172	1,169
	Sleeping Bags	2,579	1,250
	Tents	2,334	1,050
Golf Equipment	Golf Accessories	1,370	651
	Irons	1,263	626
	Putters	972	450
	Woods	1,256	606
Mountaineering Equipment	Climbing Accessories	1,655	790
	Rope	957	455
	Safety	912	473
	Tools	1,337	670
Outdoor Protection	First Aid	1,787	784
	Insect Repellents	1,678	813
	Sunscreen	1,913	961
Personal Accessories	Binoculars	2,477	1,079
	Eyewear	9,210	4,843
	Knives	2,819	1,318
	Navigation	2,948	1,440
	Watches	6,501	3,480
Total transaction records		**54,519**	**26,861**



(a) Training data

(b) Testing data

Figure 5.2- Distribution of sales transaction records by retailer country

5.3.3 Result from the first stage – Sales forecasting

Sales forecasting is generated by using the approach proposed in Chapter 4. First, the RFE algorithm is performed on the training dataset to select the most relevant predictors of sales. The dataset is also run with the Boruta algorithm for all-relevant predictor selection; however, the data size exceeds the computer memory and the running time is very long. Hence, the RFE is chosen in this case as the more cost-effective solution for the practical use. The result of the RFE is presented in Table 5.3. As seen in the table, the best predictive performance (R^2 = 0.259462) is reached with twelve predictors. The performance is reduced when adding Unemployment rate to the model. Hence, the minimal subset of the strongly relevant predictors of sales include *Selling price, Previous quarter sales, Quarter, Order method type, Retailer country, Retailer type, Product line, Product type, GDP per capita, CPI rate, Average temperature, and Average humidity.*

Table 5.3 - The RFE ranking of variable relevance based on predictive performance

Variable name	R^2	Selection decision
Quarter	0.248342	Selected
Retailer country	0.25633	Selected
Order method type	0.256976	Selected
Retailer type	0.257099	Selected
Product line	0.257398	Selected
Product type	0.257475	Selected
GDP per capita	0.257837	Selected
CPI rate	0.257987	Selected
Unemployment rate	0.258759	Unselected
Average temperature	0.258623	Selected
Average humidity	0.258619	Selected
Selling price	0.259462	Selected
Previous quarter sales	0.257994	Selected

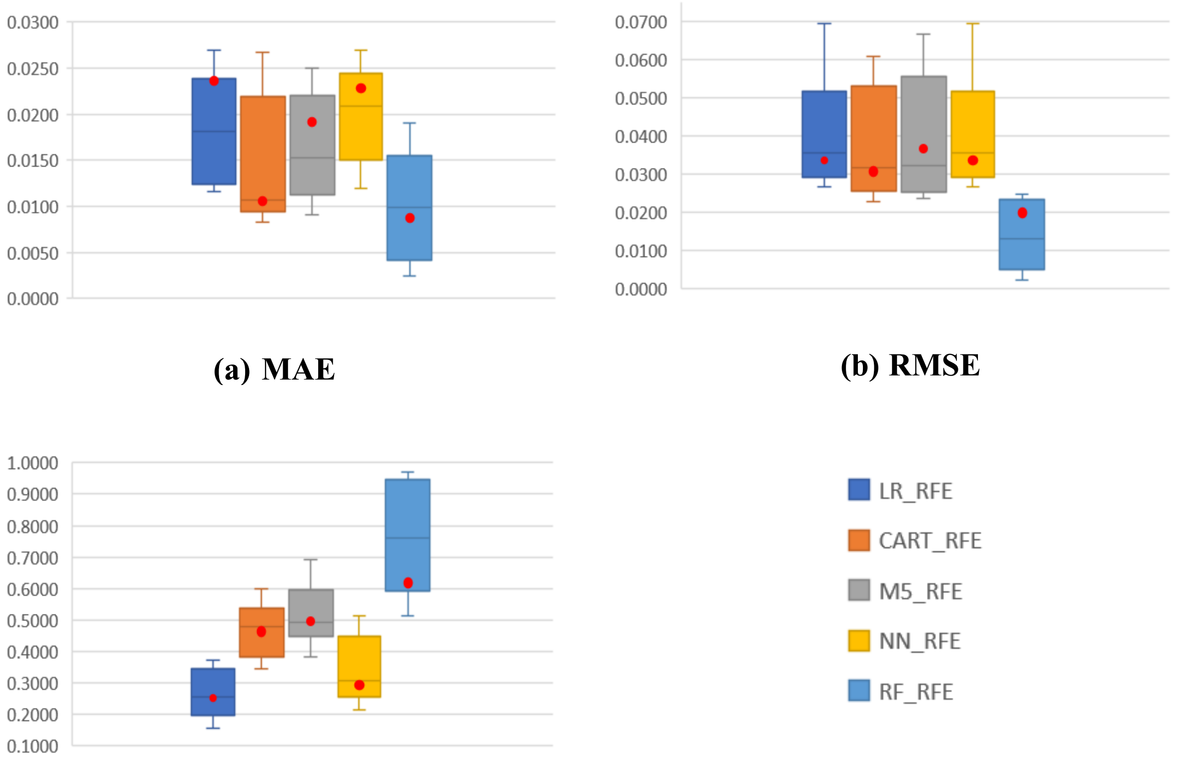
After identifying the input set with the RFE algorithm, the next step is the prediction model training. In this research, all the algorithms are fitted into the training dataset and then validated with the testing set. A ten-fold cross validation is applied throughout the modelling process in order to remove the resampling bias. To avoid bias issues, five algorithms are used to train the sales prediction model: LR, CART, M5, RF, and ANN. The model performance is assessed according to three measurements: MAE, RMSE and R^2. For a fair comparison, each model is run with its optimal hyperparameter settings and the RFE predictor set above. The result of the hyperparameter tuning is presented in Table 5.4.

Table 5.4 - Model hyperparameter tuning based on ten-fold cross-validation

Predictive algorithm	No. of models per algorithm	Tuning parameter	Candidate parameter setting	Parameter selection
CART	10	Maximum depth of the tree	{1, 2, 3, 4, 5, 6, 7, 8 , 9, 10}	11
M5	8	Pruned	{Yes, No}	No
		Smoothed	{Yes, No}	Yes
		Rules	{Yes, No}	No
RF	10	Number of variables at each split	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11}	2
ANN	20	Number of hidden units	{1, 3, 5, 8, 10, 11, 13, 15, 17, 20}	13
		Weighting decay	{0.5, 0.1}	0.1
LM	1	N/A	N/A	N/A

With the optimal parameter settings as selected in Table 5.4, the prediction results of each algorithm in the model training and testing phase are shown in Figure 5.3. RF outperforms the other predictive algorithms in both training and testing stage, with the lowest values of RMSE and MAE, as well as the highest R^2. Also, Figure 5.3 indicates no significant under-/overfitting issues since all the values of MAE, RMSE and R^2 in the testing phase (the red points) fall within their corresponding distributions in the training phase (the boxplots).

Based on the RF algorithm, the forecasted values of product sales corresponding with their prices in the testing dataset are inputted to the price optimisation model in the second stage.



(a) MAE

(b) RMSE

(c) RMSE

Figure 5.3 - Prediction results in the training (boxplots) and testing (red points) stage with ten-fold cross validation

5.3.4Result from the second stage – promotional price optimisation

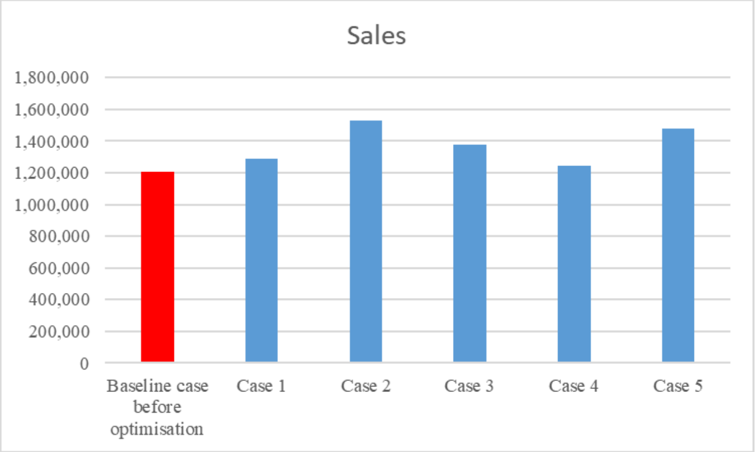
Since the promotion campaign is normally initiated in the local scale rather than global scale, the price optimisation experiment is conducted in the single country. Without losing the generalisation, the author chooses the United State as the market. The model is applied for all five product lines which include 21 types of products. The sales values forecasted based on RF model for each product types corresponding to its given unit prices from the testing data in the first stage are used as inputs in the price optimisation model. Among the five product lines, the outdoor protection line has the smallest number of product types (03 types). Therefore, the author assumes the maximum number of product types with discounted prices in each product line (max_j) is 3.

Regarding to the constraint (5.4), although the user-defined maximum revenue loss that the retailer is willing to accept for the promotion campaign (cap) is not given, it is assumed that the price promotion expenditure may take up from 10% to 50% of the expected revenue before optimisation. A sensitivity analysis on this parameter is executed to help find the bound of the optimal solution. The result of the objective value (sales) and the estimated revenue corresponding to the optimal price is summarised in Table 5.5.

Table 5.5 - Promotion cap value

	Fraction of promotion investment over the expected revenue before optimisation ($65,258,590)	**Promotion cap value (cap)**
Case 1	10%	6,525,859
Case 2	20%	13,051,718
Case 3	30%	19,577,577
Case 4	40%	26,103,436
Case 5	50%	32,629,295

The result of the optimisation model in different cases is presented in Figure 5.4. Generally, the proposed promotional price optimisation is able to find the optimal solution that increase the sales and revenue as in case 1, case 2, and case 3. Among all cases, case 2 with the 20% promotion cap ($13,051,718) seems to be the most favourable solution as it leads to the most increase in both sales (by 26%) and revenue (by 35%). On the contrary, case 5 is the most expensive case (the highest promotion expenditure), but interestingly, provides with the most undesirable solution where there is a revenue loss by about 7% in exechange of just over 22% sales increase. This is an example of excessive price discounts that would have an inverse effect on the profitability.



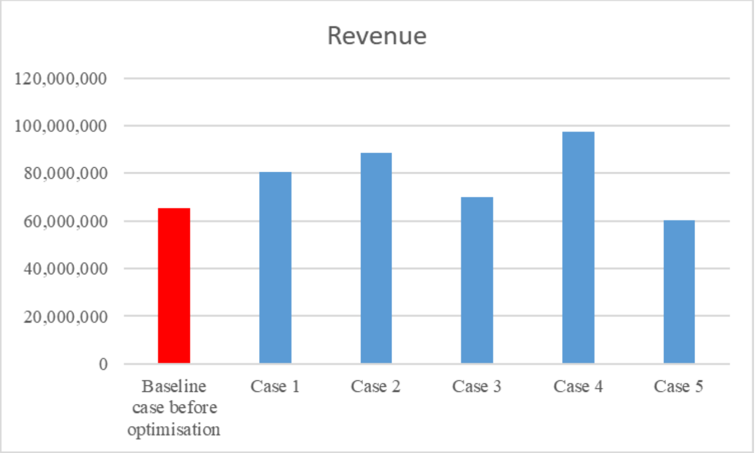


Figure 5.4 – Sales and Revenues corresponding to the optimal prices

The optimisation model can also support the decision makers in a way that identifies what product types should be put on offer in each product line and at what price they should be set. Based on case 2, the optimal promotional plan is described in Table 5.6.

Table 5.6 - Optimal price-promotion plan based on case 2

Product Family	Product type	Is it on discount?	Suggested Price
Mountaineering Equipment	Ropes	No	546.44
	Tools	Yes	58.79
	Safety	Yes	61.75
	Climbing Accessories	Yes	17.87
Personal Accessories	Binoculars	No	173.65
	Knives	Yes	88.07
	Eyewear	Yes	68.07
	Watches	Yes	73
	Navigation	No	358
Golf Equipment	Woods	Yes	1291.73
	Golf Accessories	Yes	5.36
	Putters	Yes	55.43
	Irons	No	882.1
Camping Equipment	Tents	No	815.24
	Cooking Gear	Yes	3.57
	Packs	No	437.49
	Sleeping Bags	Yes	95.48
	Lanterns	Yes	31.55
Outdoor Protection	Insect Repellents	Yes	5.96
	Sunscreen	Yes	4.96
	First Aid	Yes	15.62

5.4 Summary

This chapter develops a prescriptive promotional price optimisation empowered based on the data-driven sales forecasting. In particular, the forecasted values of the location-specific product sales are first produced by using the ML-based predictive analytics approach proposed in Chapter 4, which involves two stages. Then, the forecasted sales with their corresponding prices are used as inputs in the MILP-based price optimisation model of which the aim is to find the optimal promotional pricing plan that maximises the total sales under some business constraints. For validation, the proposed approach is applied to the real-world database of 81,380 sales transactions. The result shows that the proposed prescriptive approach is able to provide the optimal promotional pricing solution that could increase the sales by 26% and revenue by 35%. However, one of the major limitations of this model is that the MILP-based optimisation would be the most efficient to the small-to-medium promotional problem, meaning the single-period promotional planning at the regional scale, rather than global scale. While the proposed sales prediction model is scalable, future research would overcome the size limitation by using the heuristic-based optimisation approach that can effective handle the large-scale optimisation problem in the real-world business system.

Chapter 6 Data-driven prescriptive optimisation of large-scale dry port network design

Chapter 5 proposed a way to develop prescriptive analytics which is empowered by predictive analytics. Although it can generate the proactive decision making for higher business value, its main limitation is about the problem size that can be handled at its second stage of optimisation. This is because the optimisation is built based on the MILP formulation. However, the practical value of MILP-based models is still rather limited, especially when dealing with a real-world system which inherently involves millions of integer decision variables. It can easily make the MILP model computationally intractable (Fischetti et al., 2017).

Therefore, this chapter introduces a new way for prescriptive analytics empowered by descriptive analytics (association rule mining), complex network theory and heuristic-based optimisation. The main advantage of this approach is its scalability to effectively deal with the complex, large-scale system in practice. The model is validated using the real case study of dry-port location optimisation in Mainland China under the context of Belt and Road Initiatives.[3]

The chapter structure is the following. Section 6.1 introduces research problems along with research aim and objectives set out for this research. Section 6.2 reviews the related literature on dry port locations. Section 6.3 describes the proposed methodology. Section 6.4 explains the model validation from the case setting, result analysis, discussion, and robust checking. The conclusion and future research directions are in section 6.5.

[3] Chapter 6 has currently been under reviewed with minor changes by the journal. See the reference:

Nguyen, T. V., Zhang, J., Zhou, L., Meng, M. and He, Y. (2019). A data-driven optimization of large-scale dry port locations using the hybrid approach of data mining and complex network theory. *Transportation Research Part E: Logistics and Transportation Review*. (Under-reviewed with minor changes).

6.1 Introduction

With the rapid development of globalization and international trade, intercontinental freight transport has experienced a fast-paced growth rate of 9.3% per year, from just under 85 million twenty-foot equivalent units (TEUs) in 1990 to about 651 million TEUs in 2013 (Lee and Song, 2017). Nevertheless, as container flows continue to rise steeply, many seaports have been confronted with the problem of severe congestion in terminals and bottlenecks in the inland transportation system (Chang et al., 2015). Under such circumstances, dry ports have been increasingly implemented as an effective logistics solution to sustain seaport competitiveness and improve the efficiency of the freight transportation chain as a whole (Roso and Lumsden, 2010).

By definition, dry ports are inland intermodal terminals connected directly to one or several seaports by high-capacity transport modes, preferably railways, where shippers and carriers can drop off and/or pick up their containers directly as if going to seaports (Crainic et al., 2015). In general, dry ports provide almost all services offered at a seaport, such as customs clearance, storage, maintenance and repair of empty containers, tax payments, and other value-added logistics activities. By transferring these services to the hinterland, dry ports can help ease many pressures and constraints faced by seaports, such as alleviating congestion at terminals and surrounding areas, increasing berth throughputs, and improving inland accessibility as well as offering better services to shippers and transport operators (Roso et al., 2009). For the public sector, the implementation of dry ports benefits the entire inland transportation system and the general ecological environment by shifting freight flows from roads to intermodal solutions favoring rail transport, which is more environmentally friendly and cost-effective (Henttu and Hilmola, 2011). For the private sector, dry ports provide shippers, especially those far away from coastal ports, with direct access to international logistics services at lower costs and higher efficiency (Roso and Lumsden, 2010).

Operating as a consolidation point and logistics hub in the broader transport network, the success of a dry port is thus critically dependent on its location advantage (Lättilä et al., 2015). A well-selected location can help dry ports attract adequate freight volumes from inland shippers, attaining economies of scale with full train services to seaports (Roso et al., 2009).

Conversely, poorly planned dry ports can result in overcapacity, facility redundancy, a low efficiency and utilization rate, and threatening returns on investment. More importantly, once a dry port is built, it is almost impossible to relocate, because of the heavy capital investment involved and the location-bound and sunk cost nature, as Chang et al. (2015) explain. Therefore, it is imperative to optimise the location and coverage area of dry ports at an early stage of their development.

Recent research has made good progress in applying most of the traditional modelling approaches from facility and hub location theory to dry port developments (Chang et al., 2015; Witte et al., 2019). However, there is still much room for improvement. This research aims to address two important research gaps in dry port location literature.

The first research gap is that previous studies commonly adopt the multi-criteria decision-making (MCDM) approach for dry port locations to explicitly assess a wide range of conflicting objectives from various stakeholders (Canh and Notteboom, 2016; Ka, 2011; Ng and Gujar, 2009). However, most of these focus on the influencing factors supported by domestic stakeholders, such as port controllers, shippers, road/rail operators, and local governments, while the role of international customers is largely overlooked. Even though international customers do not make direct decisions about dry port locations, their market power in international trading still has a significant impact on whether or not particular dry ports attract sufficient demand to thrive and achieve self-sustainability. It is also noteworthy that the routing choice of shippers varies greatly based on the origin and destination of local suppliers and international customers as well as on order characteristics such as lead time and value. The failure to capture these types of dynamic features in a dry port location analysis will lead to suboptimal solutions, especially when international purchasing behavior changes in the future. Hence, from the microeconomic and business perspective, it would be more sensible and safe to locate the dry ports in inland regions where local suppliers have the most capability of attracting and meeting international customer demand. Since e-commerce is the main driver of international trading, mining its detail-rich, individual-level transaction data can effectively capture the dynamic behavior of international purchasing patterns. This can be done by generating an origin-destination matrix to represent the demand flow between local suppliers and international buyers. From this, the optimal location of dry ports can be determined in a way that maximizes the passing flow of international purchasing. The origin-destination matrix as a demand representation approach has been effectively used in several big data-enabled transportation studies (Cui et al., 2016; Shan and Zhu, 2015; Toole et al., 2015).

The second research gap is that most current dry port optimization models are developed based on the classical mixed-integer programming (MIP) approach (see, e.g., Ng and Gujar 2009, C. Wang et al. 2018, Wei and Sheng 2017). Despite its huge success in the literature in optimizing location problems, the practical value of MIP-based models is still rather limited, especially when dealing with a complex, large-scale system (Fischetti et al., 2017). This is because (1) the real-life logistics network, which inherently involves millions of integer decision variables, can easily make the MIP model computationally intractable, and (2) the MIP model is typically developed based on explicitly predefined network structures and simplifying assumptions, thereby considerably confining the true discovery of the optimal dry port location (Zheng et al., 2018). Therefore, there is still an absence of an optimization method that can effectively provide practical solutions for the large-scale dry port location problem.

Aiming to address both research gaps above, this research proposes a two-stage approach, called the Association Rule Mining with Eigenvector Centrality - Gravity-based Community Structure (ARMEC-GCS), a data-driven optimization for the large-scale dry port location and service area identification problem. In the first stage, we mine a large amount of international transaction data using the ARMEC model to weight the importance of inland regions based on the microeconomic and business perspectives of international customers. In the second stage, the weighting score is then integrated with other factors from macroeconomic (i.e., inland region’s foreign trades) and geographic (i.e., spatial distances between inland regions) perspectives in the GCS algorithm to optimise the location and coverage area of dry ports.

In this study, the ARMEC-GCS approach is validated using the real problem of dry port development currently faced by China in the context of its biggest megaproject, the Belt and Road Initiatives (BRI). The scalability and practical value of the proposed method is assessed in the nationwide dry port network composed of 95,481 edges between 309 city nodes in Mainland China. For the ARMEC model, the author uses a real-world Alibaba database recording 25,643 international transactions.

The contributions of this study are twofold: theoretical and practical. For the theoretical contribution, this is a pioneering study applying the data-driven approach for the large-scale dry port location optimization problem. By mining a large amount of individual-level transaction data in international trading, the location preference from international customers' perspectives, which have been unknown and neglected in prior literature, is discovered and taken into account in the decision-making process of dry port development. Also, as compared to classical methods in location theory, such as MCDM and MIP, the novelty of the proposed method is that its stages both for location importance ranking and location optimization are nonparametric and data-driven without prior assumptions made on the variable distribution. As such, the location advantage of each inland region can be truly explored in nature by letting the data speak for itself. Most importantly, this study opens up considerable opportunities to expedite the research progress and the practicability of location theory in the era of Industry 4.0 by adopting new modelling techniques from two emergent domains that have been widely used to study many real-world systems: data mining (also machine learning) and complex network theory. In this regard, a large variety of real-world big data sources (e.g., Alibaba, Amazon, and eBay) can also be leveraged for new location criteria. The research also helps promote synergies between operation research and data mining – a new, important research stream, as suggested by Corne et al. (2012).

Regarding the practical contribution, our new effective approach is able to produce a realistic and applicable dry port location solution covering the large-scale area of Mainland China. In particular, the optimal solution is derived from multiple decision-making perspectives (i.e., macroeconomic, microeconomic, and geographical), which in turn increases the possibility of its acceptance by various groups of stakeholders and of obtaining funds from the BRI investment, as this solution is practically in line with the market-based principle of the BRI, holding that although the initiative is a policy proposal, its execution must make commercial sense (Huang, 2016).

6.2 Review of dry port location studies

The current literature on dry port location analysis is summarized in Table 6.1. In general, research on dry port locations can be classified into two fundamental design perspectives: microeconomic and macroeconomic. This classification is related not only to the scope of the problem but also to the modeling method adopted.

In the microeconomic perspective, the designer makes the choice of dry port locations based on the economic benefits to be gained from the improved performance of the transportation and supply chain operations. For example, Feng et al. (2013) optimise the location and allocation of the regional seaport and dry port system with the aim of minimizing the sum of the transportation, dry port set-up, and maintenance costs. The dry port location in C. Wang et al. (2018) is selected, taking into consideration the transportation cost and the cost of opening/closing new/existing facilities. Wei and Sheng (2017) and Ng and Gujar (2009) also choose cost savings in logistics as the primary objective in their dry port location models. All these studies formulate the location optimization problem as a compact MIP model, where the optimal dry ports are selected only from a fixed set of candidate locations given in advance. Another concern is that the optimal solution may only hold true to the specific network topologies used for its model development. As a result, these simplifying assumptions seem to constrain the discovery of the truly optimal location and the practical application of the findings (Zheng et al., 2018).

On the other hand, many researchers take a broader macroeconomic perspective in which dry port locations are considered a multi-criteria decision, allowing conflicting objectives from various stakeholders to be taken into account. For instance, transportation condition, local policy environment, and regional economic development are among the common evaluation indicators for dry port locations (Canh and Notteboom, 2016; Chang et al., 2015; Ka, 2011; Komchornrit, 2017; Li et al., 2011; Wei et al., 2018). Most traditional multi-attribute methods have been successfully adapted to dry port locations, including the analytical hierarchy process (AHP) (Abbasi and Pishvaee, 2018; Ka, 2011), MCDM (Canh and Notteboom, 2016; Komchornrit, 2017), and fuzzy clustering (Chang et al., 2015; Li et al., 2011). However, one of the major drawbacks of these methods is that the weight ranking and decision rules of the location criteria are assessed according to human perception and experience, which are more or less biased and subjective and are difficult to quantify accurately (Canh and Notteboom, 2016). Another common concern is that the locations derived from the multi-attribute decision making are typically optimal at the macro level only, while from a microeconomic and operational perspective, there is no guarantee they would be able to attract sufficient demand from shippers to stay economically viable (Chang et al., 2015).

Table 6.1 - Summary of literature on dry port location problem

	Location selection perspective		Problem size*		Method
	Micro-economic	Macro-economic	Small-scale	Large-scale	
Feng et al. (2013)	✓		✓		Genetic Algorithm
C. Wang et al. (2018)	✓		✓		MIP
Wei and Sheng (2017)	✓		✓		MIP
Ng and Gujar (2009)	✓		✓		Spatial analysis; MIP
Komchornrit (2017)		✓	✓		MCDM
Li et al. (2011)		✓	✓		Fuzzy Clustering
Canh and Notteboom (2016)	✓	✓	✓		MCDM
Ka (2011)	✓	✓	✓		AHP; MCDM
Chang et al. (2015)	✓	✓	✓		FCM; MIP
Wei et al. (2018)		✓		✓	PCA; Gravity model
Abbasi and Pishvaee (2018)	✓	✓		✓	AHP; MIP
This research	✓	✓		✓	Data mining; Complex Network

** "Problem size" refers to the size of the studied dry port network, which is classified as small-scale if the study focuses on the city- or regional-level network and as large-scale if the focus is on the nationwide network.*

Some researchers are also attempting to adopt both the microeconomic and macroeconomic perspectives, to complement the way they limit each other, by developing a two-stage dry port location optimization approach. As such, a set of candidate locations is first selected using the multi-criteria model at the macro level. Then the MIP model is performed to select from the candidate set the final dry port location that can optimise the performance of the logistics network at the microeconomic and operational levels (Abbasi and Pishvaee, 2018; Chang et al., 2015).

Regarding problem size, when using a conventional location modelling approach such as MCDM and MIP, most existing models for dry port location can only address the small-scale optimization problem specific to the city- and regional-level transportation systems. Thus, the large-scale dry port location problem at the nationwide level has been studied much less. In fact, the author only found two papers in the current literature that discuss national dry port development (Abbasi and Pishvaee, 2018; Wei et al., 2018). However, the optimal locations they obtained still suffered from being highly subjective and biased, due to the use of MCDM for the location criteria ranking, as explained above.

In summary, the literature review reveals the absence of a method that can effectively and unbiasedly optimise the large-scale dry port location problem, taking into account both macro- and microeconomic design perspectives. Hence, the proposed data-driven approach, the ARMEC-GCS, which combines nonparametric, scalable algorithms from the data mining domain and complex network theory, can address the gap effectively.

6.3 Methodology

In this research, the author proposes a two-stage ARMEC-GCS approach to optimise the location and service area of dry ports in a large-scale network from both the micro- and macroeconomic-level perspectives. In the first stage, the author constructs a real-world transactional database for international purchasing records and apply association rule mining (ARM) to extract a set of association rules where international demand patterns are considered the antecedents and supplying locations are the associated consequences. Then the antecedents and consequences of all association rules are visualized as a network of international purchasing, in which the influence of each supplying location node can be measured by the eigenvector centrality (EC). In the second stage, the dry port network is regarded as a complex network where inland regions are represented by nodes and the relationship between regions can be represented by edges. In such a network, the author applies a gravity-based community structure (GCS) algorithm to seek optimal dry port locations and identify the service area covered by each. In the GCS algorithm, the logistics-connecting relationship between nodes (i.e., edge weights) is measured by an extended gravity model, taking as inputs the EC scores from the first stage as well as foreign trade values and spatial distances between nodes.

The overview of the proposed approach is shown in Figure 6.1, while the demonstration of each phase is detailed in the following subsections.

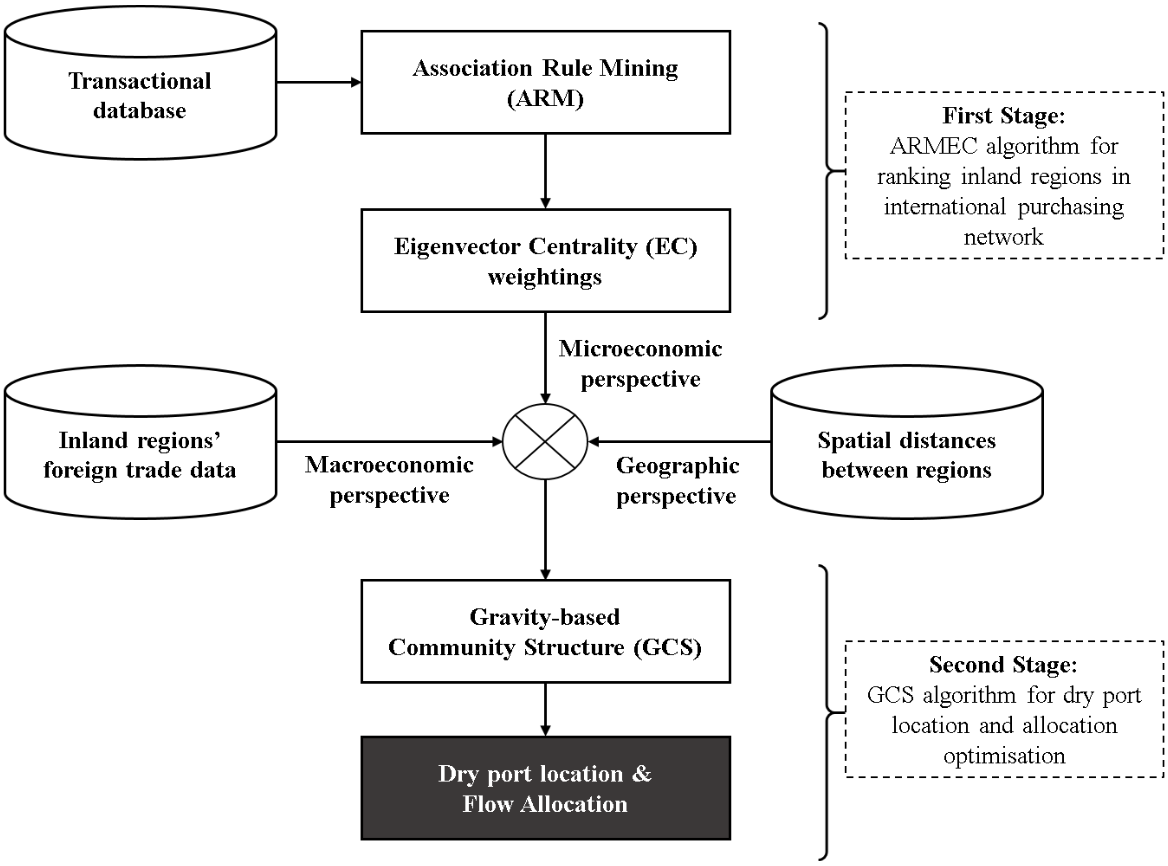

Transactional database
Association Rule Mining (ARM)
First Stage:
ARMEC algorithm for ranking inland regions in international purchasing network
Eigenvector Centrality (EC) weightings
Microeconomic perspective
Inland regions' foreign trade data
Macroeconomic perspective
Geographic perspective
Spatial distances between regions
Gravity-based Community Structure (GCS)
Second Stage:
GCS algorithm for dry port location and allocation optimisation
Dry port location & Flow Allocation

Figure 6.1 - The ARMEC-GCS approach

(Source: The author)

6.3.1 Constructing international transaction database

To identify optimal inland cities for locating dry ports based on their trading attractiveness to international customer demand, the author constructs a large-scale international transaction database recording all demand and supply information, for example, product type, transaction value, buyer location, supplier location, order date, supplier's company size, reputation, production capacity, trading capability, etc. Nevertheless, the main focus of the present study is to discover association rules between international demand patterns and supplying locations; therefore, the author only selects buyer- and product-related attributes. In particular, the international demand pattern is represented by a matrix of $D(x_i, y_i, z_i)$, in which each matrix element accommodates a key feature of the i^{th} transaction, including the buyer location (x_i), the production lead time (y_i), and the transaction value (z_i). International demand patterns are distinguished by the unique combinations of these three features. The demand matrix, together with the supplier location (s_i), formulate the transactional databased used for the ARM model. A sample of the international transactional database can be seen in Table 6.2.

Table 6.2 - Example of the international transactional database

Supplier ID	Transaction ID	Buyer location $(\boldsymbol{x})$	Lead time $(\boldsymbol{y})$	Transaction value $(\boldsymbol{z})$	Supplier location $(\boldsymbol{s})$
A1	TID1	Poland	7 days	Low	Nanjing
A1	TID2	India	30 days	Very high	Nanjing
A2	TID4	Romania	60 days	Medium	Foshan
A2	TID5	Finland	20 days	Low	Foshan
….	…	….	…	….	…

6.3.2 Stage 1: ARMEC algorithm

6.3.2.1 Association rule mining

Data mining is the process of applying a wide range of machine learning and statistical techniques in order to extract previously unknown patterns for better decision making (Corne et al., 2012). ARM is among the most versatile and widely used data mining techniques (Nguyen et al., 2018b). It is the method of finding frequent patterns, associations, co-occurrences, or causalities between a complex set of attributes in big data (Ting et al., 2014). A prime example of ARM applications is market basket analysis, which aims to extract from large-scale transactional databases a set of association rules about which products customers frequently bought together. Such rules have been well-adapted to support various decision making, for instance, new product development (Bae and Kim, 2011), logistics quality control (Ting et al., 2014), and fraud detection in procurement management (Ghedini Ralha and Sarmento Silva, 2012) . ARM has also been effectively used to optimise location-related problems, such as shelf-space allocation (Tsai and Huang, 2015), storage assignments (Chiang et al., 2011), and logistics scheduling (Lee, 2016), which makes it relevant to the studied problem of dry port location.

The output of the ARM is a set of association rules that can be expressed in the format {A} => {B}, where A and B refer, respectively, to the antecedent and consequence part of the rule. In this study, the ARM aims to evaluate the supplying capability and trading attractiveness of inland regions from the business perspective of international customers. Thus, the author only focuses on association rules for which the antecedent (A) is the set of international demand patterns and the consequence (B) is the set of suggested supplying locations.

There are many measures of rule strength or importance, as explained in De La Iglesia et al. (2006). In this research, the author uses the most common ones, namely *support* and *confidence*. The rule support refers to the probability that both the antecedent and consequent occur together, while the rule confidence is the conditional probability that the consequent occurred based on the occurrence of the antecedent (Padmanabhan and Tuzhilin, 2003). While the support implies the coverage (or frequency) of the rule in the transaction database, the confidence indicates the rule strength (or reliability) (Witten and Frank, 2011). Typically, a rule is considered as important and interesting if it satisfies both the minimum support and minimum confidence thresholds predefined by domain experts. The mathematical expression of support and confidence is as follows:

$$Support = P(A \cap B) = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions} \tag{6.1}$$

$$Confidence = \frac{P(A \cap B)}{P(A)} = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} \tag{6.2}$$

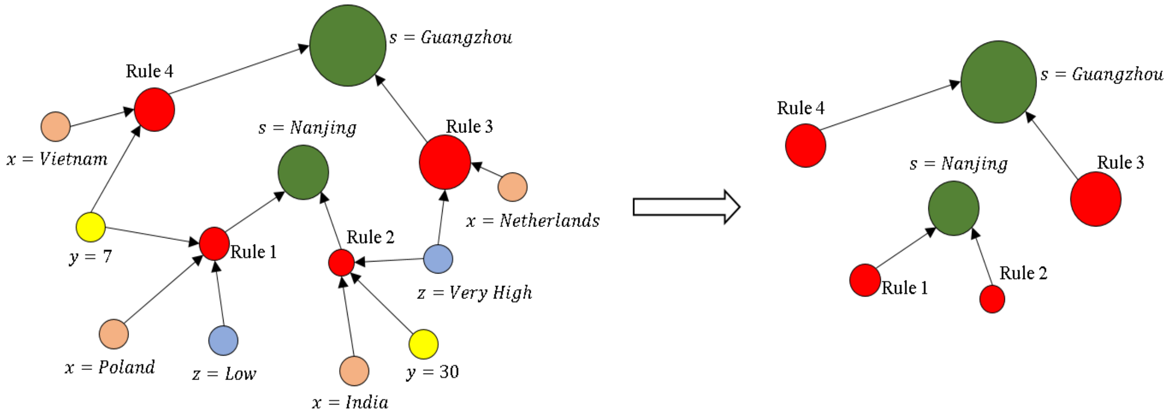
Given the fact that the number of rules grows exponentially, which makes the brute-force approach infeasible, this research thus adopts one of the most popular ARM algorithms, called Apriori (Ghedini Ralha and Sarmento Silva, 2012). The algorithm involves two phases. In the first phase, it performs a breadth-first search to generate a large set of candidate item sets from which frequent item sets are identified. The principle is that an item set is considered a frequent item set if all of its subsets have support higher than the predefined minimum support threshold.

In that way, the searching space can be reduced significantly, since only frequent item sets are used to generate association rules in the second phrase. Similarly, only rules that have confidence higher than the predefined minimum confidence threshold are considered interesting and worth further analysis.

6.3.2.2 Eigenvector centrality in complex network

Parallel to rapid progress in studying big data analytics, another emerging research stream is big data visualization, which involves multiple techniques to make the result of big data analytics more understandable, accessible, and useable for timely data-driven decision making (Nguyen et al., 2018b). Among different visualization techniques, complex network analysis has been proven one of the most scalable techniques for dealing with large, complex data. Unlike classical network theory, the complex network focuses primarily on studying the nontrivial topological patterns that are neither uniformly ordered nor random (Rubinov and Sporns, 2010). Since such complex patterns are inherently linked to most real-world systems, the method has gained much attention from researchers from a wide range of fields, such as biology (Rubinov and Sporns, 2010), transportation (Saberi et al., 2017), and social networks (Verma et al., 2018). There are a number of measures to describe the structural properties of a complex network. In this research, the author uses two fundamental measures: EC for the ARMEC model and community structure for the GCS model (see section 6.3.3).

After generating a set of association rules by using the Apriori algorithm described in section 6.3.2.1, the next step in the ARMEC model is to develop an international purchasing network in which all objects and relationships among the association rules are represented as nodes and edges. In such a network, nodes include association rules, their associated antecedents (i.e., international demand patterns), and their associated consequences (i.e., supplying locations). Causal relationships among the association rules are illustrated by directed edges. An example of this network can be seen in Figure 6.2a. However, within the scope of this study, the author focuses particularly on the nodes representing the supplying locations; therefore, the network excludes nodes representing demand patterns, as seen in Figure 6.2b. In network (b), the size of the red node represents the strength of the association rule measured by its confidence value, whereas the size of the green node indicates the centrality of the supplying location in the network.



(a) Full network

(b) Simplified network

Figure 6.2 - Example of international purchasing network used in this study

Red node - Association rule ID; Green node - Supplying location; Orange node - Buyer location; Yellow nodes - Product type based on production lead time (days); Blue nodes - Transaction value

In network analysis, node centrality refers to the importance of a node in the network. There are various indices to measure node centrality, including degree, closeness, eigenvector, clustering coefficient, betweenness, and information index (Wang, Li, et al., 2018). In this research, the author uses EC to measure the importance of supplying location nodes in the international purchasing network (Figure 6.2b). Ghanbari et al. (2018) define the EC as below.

Let $G = (V, E)$ be the network G containing a set of nodes V and a set of edges E. The network can be represented through its adjacency matrix $A = \{A_{ij}\}$, where A_{ij} is a binary variable that takes 1 if an edge exists between node v_i and node v_j and takes 0 otherwise. The EC of a node is determined by taking into account both the number of its connected neighbors and the importance of each neighbor. Hence, the EC k_i of node v_i is proportional to the sum of the centralities of its connected neighbors. As such:

$$k_i = \frac{1}{\lambda}\sum_{1}^{n} A_{ij}x_j \qquad (6.3)$$

where n is number of neighbors linked with node v_i, x_j is the eigenvector value of the neighbor node v_j, and λ is the largest eigenvector value in the adjacency matrix A.

6.3.3 Stage 2: GCS algorithm

The dry port-based inland transportation system could be regarded as a complex network where the ports could be represented by nodes and the relationship among ports could be represented by edges. To determine dry port locations and their coverage areas in such a network, we adopt the concept of community structure from complex network theory. Community structure (so-called clusters or modules) is a common phenomenon in many real-world networks, referring to partition of a network into groups (or communities) of nodes which are densely connected within the groups and sparser connected with nodes in other groups (Costa, 2015). Several studies have been published using community structure theory in the transportation and logistics research area such as cargo ship movement analysis (Kaluza et al., 2010), global logistic network design (Sun et al., 2012) and global hub location optimization (Zheng et al., 2018). In general, it is feasible to use community structure theory to detect port relationships at a large-scale network level.

A range of approaches have been developed to detect the community structure in complex networks, for example, spectral-based, clustered-based, and modularity-based algorithms (Zhou et al., 2018). Among these, the modularity-based algorithm has been widely applied in large-scale networks, due to its fast, efficient computation (Clauset et al., 2004). Modularity is a quality function to measure whether a particular partition of the network into communities is good, in the sense that there is a high density of edges within communities and only sparse connections between them. Newman and Girvan (2004) define modularity (Q) as follows:

$$Q = \sum_i (e_{ii} - a_i{}^2) \tag{6.4}$$

where e_{ii} equals to the fraction of edges that connect vertices within community i. It is the main diagonal elements of the symmetric matrix $E = \{e_{ij}\}$, where element e_{ij} is the fraction of edges in the network that connect vertices in community i to vertices in community j. The mathematical expression of e_{ij} is given by Clauset et al. (2004) as follows:

$$e_{ij} = \frac{1}{2m}\sum\nolimits_{uv} A_{uv}\ \delta(c_u, i)\delta(c_v, j) \tag{6.5}$$

where A_{uv} is an element of the adjacency matrix, which takes 1 if vertex u and vertex v are connected, and 0 otherwise; m is the total number of edges in the network, measured by $\frac{1}{2}\sum_{uv} A_{uv}$. If vertex u belongs to community i, then $\delta(c_u, i)$ equals to 1, and -1 otherwise. Similarly, if vertex v belongs to community j, then $\delta(c_v, j)$ equals to 1, and -1 otherwise.

Furthermore, ${a_i}^2$ in Eq (6.4) is the expected fraction of edges that connect to vertices in community i when the end of edges are connected at random. The expression of a_i is formulated in Clauset et al. (2004) as follows:

$$a_i = \frac{1}{2m}\sum_{i} d_u \delta(c_u, i) \tag{6.6}$$

Where d_u is the degree centrality of vertex u, measured by $d_u = \sum_1^n A_{uv}$.

Here, the modularity-based community detection model becomes a mixed-integer quadratic programming problem of which the objective is to find the optimal splitting point of the network to maximize the modularity in Eq. (6.4). Previous studies have addressed the modularity maximization using both exact (eg. Costa 2015) and heuristic approach (eg. Santiago and Lamb 2017). However, when dealing with large-scale, real-world facility location problems, using approximate optimization techniques such as greedy heuristic is an ideal choice to effectively search over a large feasibility space for optimal solutions (Ishfaq and Sox, 2011; Ruiz et al., 2018; Santiago and Lamb, 2017). Therefore, to optimise the location and service area of dry ports, this paper employs one of the most widely used algorithms in the modularity-based community structure theory, called the fast Newman (FN) algorithm (Newman and Girvan, 2004). It adopts an agglomerative approach to search the optimal network splitting points in a greedy manner.

However, the classical FN algorithm was developed specifically for an unweighted network, while the dry port transportation system is typically a weighted network of which edge weights indicate the logistics relationships between nodes. Hence, in this paper, we adopt the improved

FN algorithm which can also be used for the weighted network (Liu et al., 2013; Newman, 2004a; Zhang and Meng, 2019). In particular, e_{ij} in Eq. (6.5) and a_i in Eq. (6.6) are redefined as:

$$e_{ij} = \frac{1}{2w}\sum\nolimits_{uv} W_{uv}\ \delta(c_u, i)\delta(c_v, j) \tag{6.7}$$

$$a_i = \frac{1}{2w}\sum_{u} W_u\delta(c_u, i) \tag{6.8}$$

where W_{uv} is the edge weight between vertex u and vertex v; W_u is the vertex weighted degree, which equals to the summation of edge weight attaching to vertex u; and w is the summation of edge weight in the network, measured by $\frac{1}{2}\sum_{uv} W_{uv}$.

In this study, the edge weight, which indicates the logistics relationship between two locations, is measured using the gravity model. Based on Newton's universal law of gravity, the gravity model provides a realistic, applicable tool to describe and predict the interaction between objects, taking into account both their mass and spatial characteristics (Campbell and O'Kelly, 2012). The model has been widely applied to international trading networks, logistics hub locations, and in many other social science research fields (Anderson and van Wincoop, 2003; khosravi and Akbari Jokar, 2017; Zeng et al., 2017; Zhang and Meng, 2019). In this study, the gravity model is extended to measure the logistics relationships among inland regions, based on their spatial characteristics and logistic quality from both the macroeconomic and microeconomic perspectives. The extended gravity function measuring the edge weight W_{uv} between region (vertex) i and j in the dry port network is expressed as follows:

$$W_{uv} = \frac{T_u\, T_v}{{D_{uv}}^2\, (1 - Z_{uv})^2} \tag{6.9}$$

where D_{uv} is the spatial distance between regions u and region v. T_u, T_v are the logistics quality of regions u and region v from the macroeconomic perspective. Since the main function of dry ports is to improve the connectivity between inland regions and international gateways (eg. Seaports or cross-border train stations) for increased international trading, T_u, T_v can be measured by the total value of import and export trade through regions i and j, respectively. Prior literature has adopted such foreign trade values as evaluative criteria for dry port locations at the macro level (Chang et al., 2015; Li et al., 2011; Wei et al., 2018). Finally, Z_{uv} is the gravity coefficient adopted in the gravity function to represent the external force affecting the logistics interaction between two regions. As discussed above, the ARMEC model distinguishes the difference between regions by their EC scores, which weight the importance of regions in the international purchasing network. Since the EC score of a region depends critically on its associations with purchasing patterns of international customers, it can be used to represent the logistic quality of regions from microeconomic and business perspectives. Thus, the gravity coefficient Z_{uv} can be calculated by:

$$Z_{uv} = k_u k_v \tag{6.10}$$

where k_u, k_v are the EC scores of regions u and region v, respectively, obtained by Eq. (6.3) in the ARMEC at stage 1.

From all the adjustments above, the classical FN algorithm is elaborated to fit the weighted network of dry ports in our study. We call the new algorithm the gravity-based community structure (GCS). The main steps of the GCS algorithm are as follows:

Step 1: Network initialization: Convert the studied geographical area into an unweighted network with nodes (cities) and edges.

Step 2: Converting the unweighted network into the weighted network by calculating the edge weight W_{uv} between any pair of nodes, using Eq. (6.9). In this network, each node is treated as one community.

Step 3: Community combination.

- Sequentially join any two communities together and calculate the modularity variation ΔQ. Based on (Newman, 2004b), ΔQ is computed by:

  $$\Delta Q = e_{ij} + e_{ji} - 2\, a_i a_j = 2(e_{ij} - a_i a_j) \tag{6.11}$$

  where e_{ij}, a_i, and a_j are obtained using Eq. (6.7) and Eq. (6.8).

- On the basis of the greedy algorithm, select the join that results in the maximum increase or minimum decrease in modularity. The modularity of the new communities is computed.

Step 4: Update the elements e_{ij}.

Step 5: Execute Step 3 and 4 repetitively until the whole network is merged into one community.

Step 6: The best division is selected with the highest modularity in the process. As a result, the network is split into a set of communities. In each community, the vertex with highest weight (most influential) is selected to locate a dry port hub, fed by other vertices within the same community. The weight (r_u) of vertex u is calculated as follows:

$$r_u = \sum_v W_{uv} \tag{6.12}$$

where W_{uv} is the weight of the edge having connection to vertex u, measured by Eq. (6.9).

6.4 Experiment

In this experiment, the author applies the proposed ARMEC-GCS approach to find optimal locations of dry ports and their service areas in Mainland China in the context of the BRI framework. China is chosen as the case application in this study given the fact that the country has recently initiated a number of dry port development projects as the key enabler to reach its full international trade growth potential (Wei et al., 2018; Xie et al., 2017)

6.4.1Case study: Dry port developments under China's Belt and Road initiative

In 2013, China launched the BRI to enhance the infrastructure connectivity between Asia, Europe, and Africa, laying a stronger foundation for international trade and regional economic growth (Huang, 2016). Since then, the BRI has become one of the world's largest infrastructure and investment projects in history, with the participation of 65 countries, accounting for 63% of the world population and 30% of the global gross domestic product (Sarker et al., 2018). It is estimated that the total investment in BRI projects will reach up to USD 7.4 trillion by 2030, more than 80% of which will be in infrastructure development for two mega projects: the Belt and the Road (Swiss Re Institute, 2017). The "Belt" refers to the "Silk Road Economic Belt" (SREB), comprising six international overland corridors connecting China with Central Asia, West Asia, the Middle East, and Europe; the "Road" refers to the sea routes called the "21st

Century Maritime Silk Road" (MSR), linking the South China Sea, the South Pacific Ocean, and the Indian Ocean (Chen et al., 2018). The geographical coverage of the BRI is depicted in Figure 6.3.

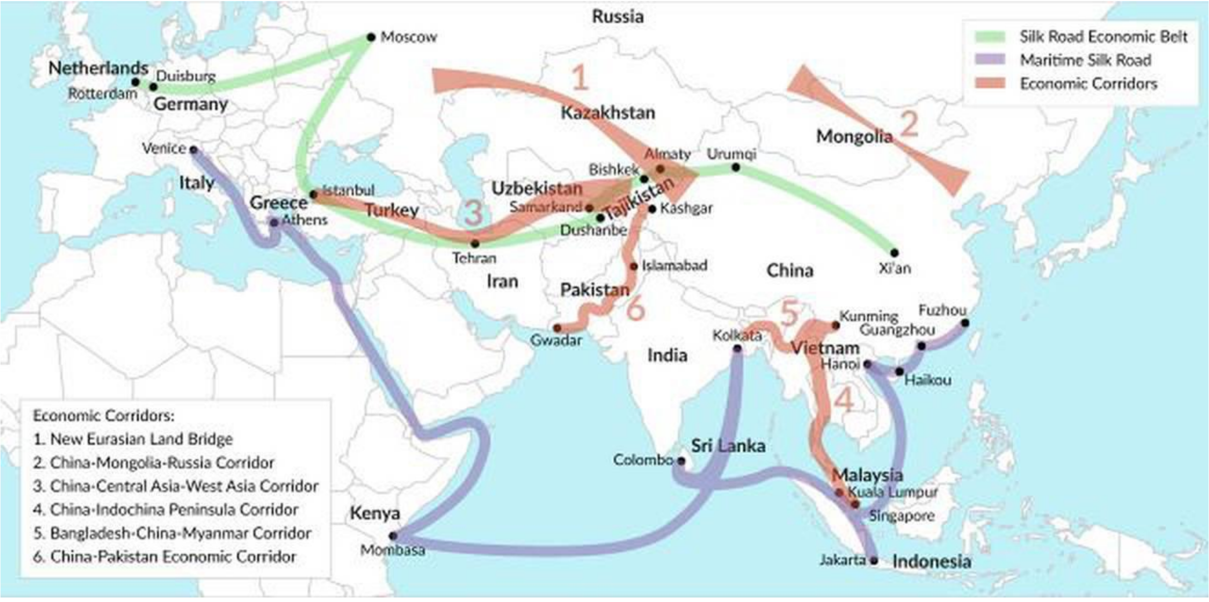


Figure 6.3 - The Belt and Road framework

A recent report by Konings (2018) claims that in the long run, the improvement in transport facility will halve overall trade costs between the BRI countries and will increase their cross-border trade by 35%–45%. Under such circumstances, using dry ports to ease congestion at port gateways and improve inland access is particularly essential to guarantee the efficiency of the entire transportation chain (Yu et al., 2018). In fact, dry ports have been set to play an integral part in the future implementation of the BRI framework, as stated by Ministry of Transport of the People's Republic of China (2017). However, the Ministry also described the current development of dry ports in Mainland China as blind constructions with a lack of unified strategic planning. Hence, this experiment aims to test whether the proposed ARMEC-GCS approach can provide a valid and applicable solution for the large-scale problem of dry port locations in China.

In particular , the author aims to find optimal dry port locations and their service areas to cover all 309 prefecture cities in Mainland China, apart from those like Qinghai, Tibet, and Guizhou Province, without a dry port operation in place (Wei et al., 2018). These studied inland cities come from 24 inland provinces, namely Sichuan, Anhui, Fujian, Gansu, Guangdong, Guangxi, Hainan, Hebei, Heilongjiang, Henan, Hubei, Hunan, Inner Mongolia, Jiangsu, Jiangxi, Jilin, Liaoning, Ningxia, Shaanxi, Shandong, Shanxi, Xinjiang, Yunnan, Zhejiang. The location problem investigating up to 95,481 edges among 309 city nodes is one of the most extensive networks in the dry port location literature, which demonstrates the real need to use scalable solution approach such as the ARMEC-GCS.

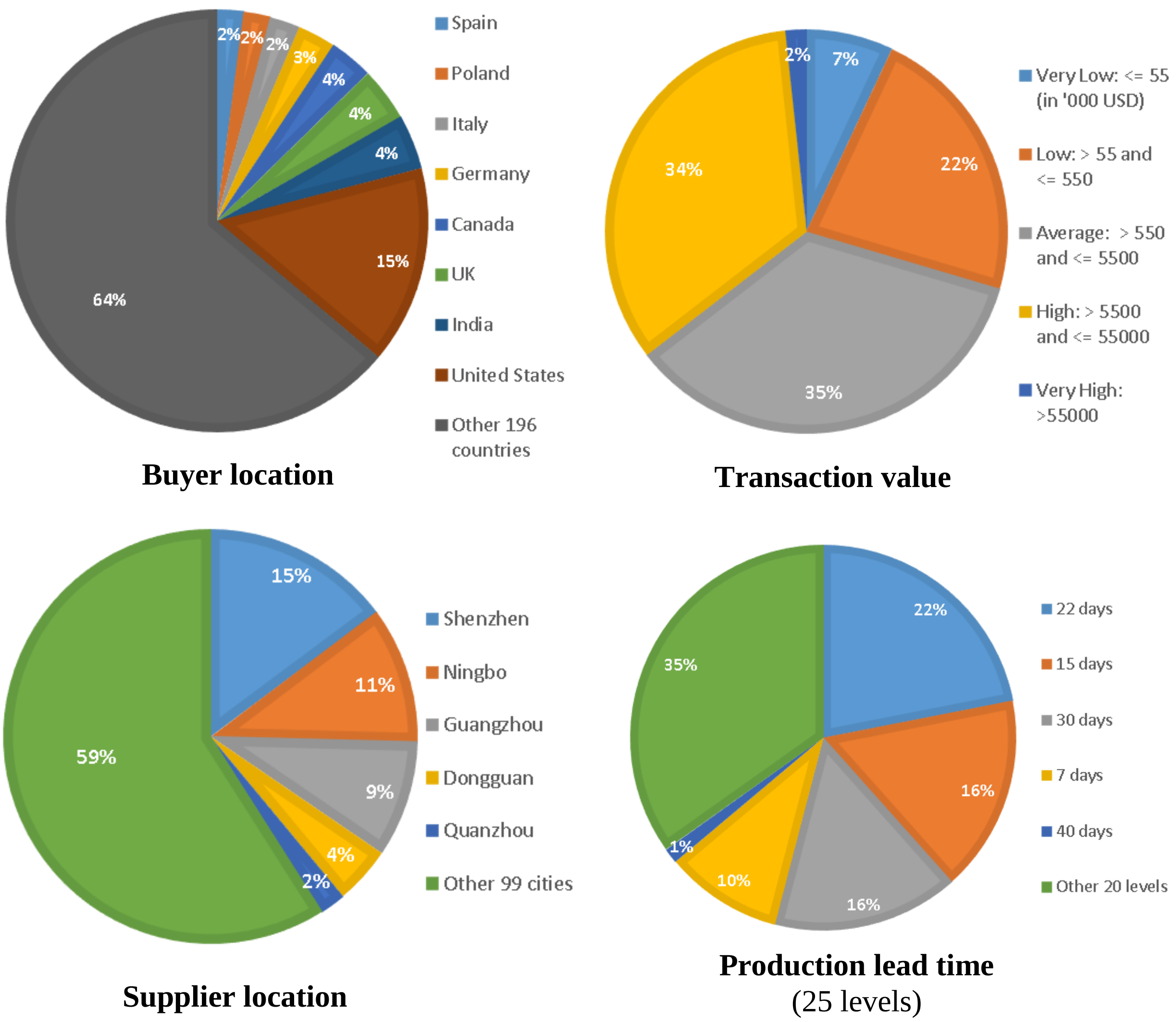
6.4.2 Data collection

6.4.2.1 Data collection for stage 1

In the first stage of the ARMEC-GCS approach which extracts insights between international demand patterns and Chinese suppliers, we construct a large transactional database from Alibaba.com. Alibaba is chosen not only because it is the world's biggest data source for business-to-business international trading, covering over 200 countries and regions, but also due to its pivotal role in the development of the Digital Silk Road as part of the BRI framework (Silin et al., 2017). In fact, Alibaba is currently developing 14 data centers around the globe, equipped with a 5G communication network, with the aim of supporting goods movement and unifying custom procedures among 10 countries along the SREB (Silin et al., 2017).

We use a web crawler to collect supplier information and sales transaction records from Alibaba.com. Since one of China's main economic interests in the BRI is to boost its inland regions towards an export-oriented economy (Huang, 2016; Wei et al., 2018), we only collect data from Chinese suppliers who provide international shipping routes across countries within the BRI projects. The transaction data we collect in this study include machinery and equipment, as they account for more than 50% of total China exports to the EU (Konings, 2018).

As a result, our crawler returns two separate datasets. The first dataset contains supplier information, while the second provides the whole transaction history of each supplier. These datasets can be joined together for data mining through the suppliers' unique IDs. After removing missing data, excluding domestic transactions and joining the two datasets, our joint dataset includes 25,643 transactions between China and international customers. Each transaction is featured by 45 attributes from the buyer and supplier. Numerous data are stored, but not all can be used to model international demand patterns. As described in section 3.1, we represent an international demand pattern through a matrix of $D(x_i, y_i, z_i)$, a compound of the transaction's buyer location (x_i), production lead time (y_i) and transaction value (z_i). The demand matrix is then mined by the ARM to find associations with the Chinese supplying location (s_i). The overall description of the Alibaba international transaction database used in this experiment can be seen in Figure 6.4.



Buyer location

Transaction value

Supplier location

Production lead time
(25 levels)

Figure 6.4 - Overall description of the Alibaba international transaction database. (25,643 total transactions)

6.4.2.2 Data collection for stage 2

In the second stage, we apply the GCS algorithm, as described in section 6.3.3, to find the optimal locations for dry ports as well as to determine the coverage area of each dry port. As explained in Eq. (6.9), the input data for the extended gravity model to measure the logistic relationships (i.e., edge weight, W_{uv}) between inland cities includes: (1) The gravity coefficient (Z_{uv}) based on the EC scores (k_u, k_v) of each city obtained from the ARMEC stage; (2) The logistics quality (T_u, T_v) of each city measured by the total import and export value obtained from its 2016 Statistical Yearbook; and (3) The spatial distance (D_{uv}) between each pair of city nodes, measured in miles, based on their longitude and latitude coordinates.

6.4.3 Experiment results and discussion

6.4.3.1 Results from the first stage - ARMEC

- **ARM results**

Based on the constructed Alibaba database, the author performs the Apriori algorithm in the R program to extract the association rules between international demand patterns (antecedent) and Chinese supplying locations (consequence). Regarding the minimum support and minimum confidence thresholds, many studies tend to set them at relatively high values to limit the number of rules generated, and decision making is derived only based on the top rules with the highest support and confidence (Ting et al., 2014). However, in order to evaluate the scalability of the proposed approach, this experiment is conducted with very low minimum support and confidence thresholds, to ensure no important rules are missed out. Since the lowest occurrence frequency for item sets in the Alibaba dataset is 0.000037, it is reasonable to set the minimum support threshold equal to 0.000037. As the transaction data in this research are sparse, the value of the minimum confidence threshold is set at its first quantile of 0.4 (40%) to avoid over-pruning informative rules while ensuring the trivial rules are excluded, as suggested by Belyi et al (2016). As a result, a total of 3,110 association rules are generated, and these international demand patterns (antecedent) are satisfied by 80 inland Chinese cities (consequence). Table 6.3 provides the statistical summary for international demand patterns within these rules. Examples of the top 10 rules with the highest confidence can be seen in Table 6.4.

Table 6.3 - Statistical description of the distribution for 3,110 rules

Antecedent size	Number of rules	Support			Confidence		
		Min	Mean	Max	Min	Mean	Max
1-itemset	81	0.00004	0.00040	0.00370	0.4	0.69775	1
2-itemset	1419	0.00004	0.00018	0.01778	0.4	0.69485	1
3-itemset	1610	0.00004	0.00004	0.00312	0.4	0.67212	1

Table 6.4 - Top 10 out of 3,110 rules sorted by confidence

Antecedent of the rule	Consequence of the rule	Support	Confidence
{ x =Niger}	=> { s = Zhongshan}	0.000039	1
{ x =Jersey}	=> { s = Zhangzhou}	0.000039	1
{ x =Indonesia, y =50}	=> { s = Tangshan}	0.000273	1
{ x =United Kingdom, y =4}	=> { s = Chengdu}	0.000117	1
{ x =Luxembourg, z = Very high }	=> { s = Quanzhou}	0.000195	1
{ x =Switzerland, y =3}	=> { s = Ningbo}	0.000078	1
{ x =Afghanistan, z = Very high }	=> { s = Foshan}	0.000156	1
{ x =Austria, y =15, z = High }	=> { s = Anqing}	0.000078	1
{ x =South Africa, y =30, z = Low}	=> { s = Jinzhou}	0.000039	1
{ x =Iceland, y =20, z = Average}	=> { s = Shantou}	0.000156	1

- **EC-based importance of Chinese cities in international purchasing network**

While the previous section determines a set of frequent rules in general, this section will demonstrate the advantage of the proposed approach, which uses a complex network to deal with the large-scale, complex relationships among these rules. In particular, all 3,110 association rules found can be visualized as a network, using popular software called Gephi. Since the main focus is on the Chinese supplier locations, Figure 6.5 displays the network that describes the relationship among the 3,110 rules (red nodes) and their associated consequences of 80 Chinese supplying cities (green nodes). The size of the red nodes represents the strength of the association rules, measured by their confidence values, whereas the size of the green nodes indicates the influence of Chinese inland cities, measured by their EC score analysis. The full list of EC scores for 80 cities is provided in Appendix.

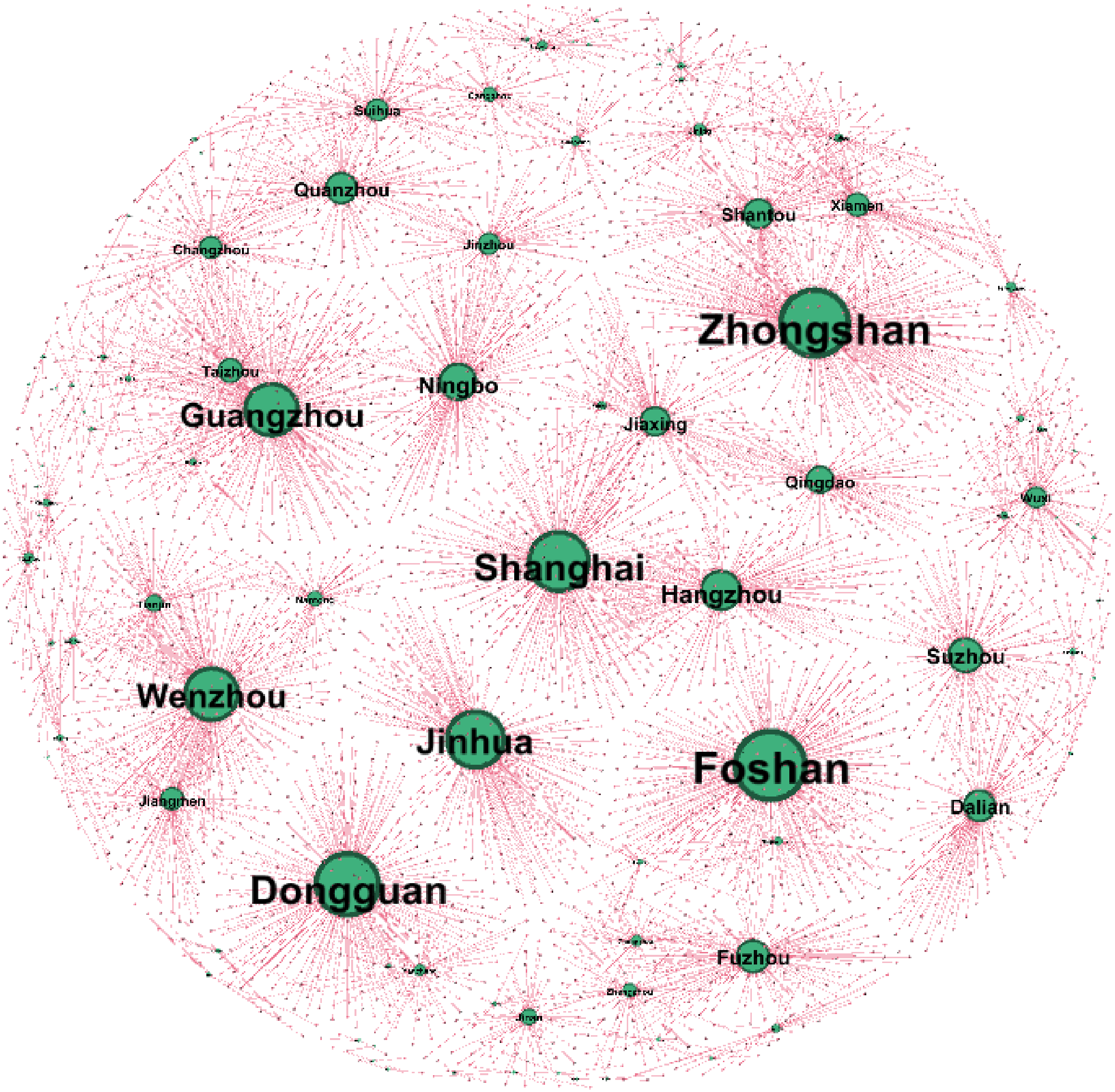


Figure 6.5 - An international purchasing network that visualises 3,110 association rules (red nodes) and their associated Chinese supplying cities (green nodes).

6.4.3.2 Results from the second stage – GCS algorithm

With the input data described in section 6.3.3, the GCS algorithm in the second stage is run on the MATLAB program. According to the final result returned from the GCS algorithm, 309 inland Chinese cities from 24 provinces are grouped into 13 communities, in which each community is served by a hub dry port with the highest degree of centrality (i.e., the most influential) in the community. The suggested dry port locations and their coverage areas are presented in Figure 6.6. For the resulting discussion, we also include in the figure the locations of major seaports and international train gateways under the BRI framework. The map codes can be seen in Tables 6.5 and 6.6.

In reality, the current dry port development in China is characterized by blind construction and a lack of strategic planning (The Ministry of Transport of the People's Republic of China, 2017). As a result, more than 100 dry ports have been built to serve the demand of over 300 inland cities. Such excessive construction could lead to overcapacity, a low utilization rate and limited returns on investment (Chang et al., 2015).

Using the ARMC-GCS approach, our results show that the Chinese inland transportation system has a strong community structure since 309 cities can be clustered cohesively into 13 communities. The average community size is quite large, which implies that the suggested dry port location in each community has strong hub functions and can attract sufficient demand to be financially viable.

The optimal dry port locations pinpointed by the ARMC-GCS approach are also closely in line with the real BRI development plan. Among 13 optimal locations, some already has the ongoing BRI dry port development projects such as Shenyang, Xi'an, Chaozhou and Xingtai, while the others currently serve as the BRI international gateways such as Beijing, Urumqi, Chengdu, Guangzhou, Suzhou, Yantai, and Xiangtan.

Moreover, the ARMC-GCS approach is also credible in terms of capturing real spatial characteristics when detecting the distinctive community structure of Community 13 (Bayingolin Mongol Autonomous Prefecture) and Community 2 (Jiayuguan and Jiuquan). Indeed, these two communities have demographic mechanisms different from other subdivisions in Xinjiang and Gansu provinces.

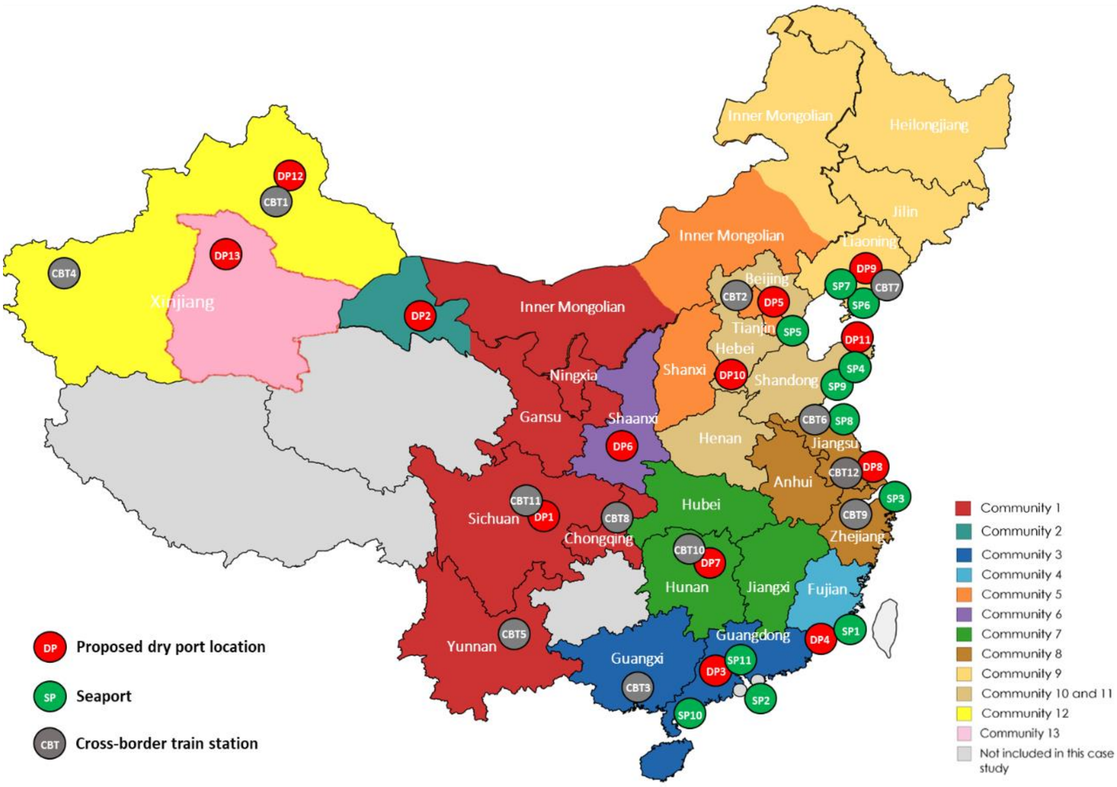


Figure 6.6 - Dry port locations and their coverage areas by the ARMEC-GCS approach

(Node codes are shown in Tables 6.5 and 6.6)

The role of each suggested dry port location in the BRI's actual development plan is highlighted in Table 6.5. Since the optimal solutions include the key transportation hubs which closely reflect the real BRI development plan, the ARMEC-GCS approach is validated.

Table 6.5 - Dry port locations with assigned communities using the ARMEC-GCS

Optimal solutions in this study			Roles of the proposed dry port locations in the BRI's actual development plan
Dry port code	Dry port location name	Number of cities allocated to the dry port	
DP1	Chengdu	56	Chengdu is the largest trade center in Western China and also the Asia's largest rail freight transport hub (Post and Parcel, 2016). One of the three key NELBEC[1] projects is Chengdu – Lodz (Poland) (Yang, Pan, et al., 2017).
DP2	Jiayuguan	2	Jiayuguan is the key transportation hub in Western China for the SREB plan. Especially, it sits one of the three key NELBEC projects, Chongqing – Duisberg (Germany), the one with numerous road and railway connections to transport goods from China to Central Asia and EU (Samaa Digital, 2017).
DP3	Guangzhou	33	Guangdong province is the key manufacturing hub having the largest export value among all Chinese provinces and municipalities (HKTDC research, 2019), while its capital city, Guangzhou, gains global recognition as the largest seaport in China and among the leading ports in the MSR (China Daily, 2018). Thus, setting up a dry port in Guangzhou to support the increasing freight traffics in the area is beneficial.
DP4	Chaozhou	14	In the implementation scheme of Guangdong's participation in the construction of the BRI, Chaozhou port is set to play supporting roles to the major seaports in the MSR like Guangzhou and Shenzen (China Daily, 2015).
DP5	Beijing	20	Beijing plays the pivotal node in both the MSR and SREB. It has direct access to Port of Tianjin which is the largest seaport in Northern China, serving 11 northern provinces and also Mongolia. It is also the starting point for one of the two major routes in the CMREC[2], namely Beijing - Tianjin - Hebei - Hohnot - Mongolia – Russia (Lehman Brown International Accountants, 2017).
DP6	Xi'an	10	Xi'an is a critical node in the BRI because it is the starting point of the New Silk Road. It also serves as transportation, trading and logistics hub connecting Northwest, Eastern, Central, and Southwest regions of China (KPMG China, 2018). Currently, there is a project to build an international dry port in Xi'an (The Ministry of Transport of the People's Republic of China, 2017).

DP7	Xiangtan	31	Xiangtan is an important node in the NELBEC. Indeed, the first China-EU train route in use was the railway starting from Xiangtan to Hamburg (Germany). Operating since 2008, the route has become the showcase for the economic advantages of the SREB-related projects (Railwaypro.com, 2017).
DP8	Suzhou	34	About 10% of all of China's exports come from Suzhou, and one of the main China-EU Silk Road route is the rail service from Suzhou to Warsaw (DHL, 2016). Suzhou also has direct connections to three major BRI international gateways in Ningbo, Jinhua and Lianyungang. Ningbo is the busiest seaport in China, and is also an intersection for both SREB and MSR (en.people.cn, 2018). Jinhua is the home of the Yiwu – Madrid international railway line - the longest railway in the world (13,000 km). Lianyungang is among the Chinese busiest seaports and the starting point of the NELBEC to Rotterdam (Sarwar, 2018).
DP9	Shenyang	42	Currently, Shenyang already has a dry port that consolidates cargoes from Anshan, Benxi and Tieling; and then, transporting by shuttle trains to Port of Dalian (Chang et al., 2015). Furthermore, Shenyang also lies on one of the two major routes in the CMREC, namely the Dailan - Shenyang - Changchun – Harbin (Lehman Brown International Accountants, 2017).
DP10	Xingtai	50	Xingtai serves as a transport hub that connects the Central China with the Eastern and Northern China. Currently, it also has a dry port partnered with Tianjin seaport (The Ministry of Transport of the People's Republic of China, 2017).
DP11	Yantai	3	Yantai is the transport hub in Eastern China's Shandong province. In 2017, it was awarded as one of the most dynamic cities in the BRI (China Daily, 2017). In 2019, it launches a new freight railway to Duisburd, Germany (Belt & Road News, 2019).
DP12	Urumqi	13	Urumqi is a key gateway in the SREB with three out of six economic corridors passing through, namely, NELBEC, CCAWAEC3, and CPEC4 (Swiss Re Institute, 2017).
DP13	Korla	1	Our model detects Community 13 due to its unique geographical position. It covers the Bayingolin autonomous prefecture for Mongol people in the southeast of Xinjiang. This is also the largest prefecture-level division in China. Setting up a dry port in its capital city, Korla, can help connect the local economy in Bayinggolin with the SREB international gateways in Urumqi and Kashgar, thereby boosting its economic growth.

[1] *New Eurasian Land Bridge Economic Corridor (NELBEC)*

[2] *China-Mongolia-Russia Economic Corridor (CMREC)*

[3] *China-Central Asia-West Asia Economic Corridor (CCAWAEC)*

[4] *China-Pakistan Economic Corridor (CPEC)*

Table 6.6 - Major international gateway ports under the BRI and suggested partnerships with dry ports based on the results of this experiment

Function	**Code**	**Actual international gateway**	**Suggested partnerships with hub dry ports obtained from this study**
Seaport (SP)	SP1	Xiamen	DP4, DP7
	SP2	Shenzen	DP3
	SP3	Ningbo	DP8
	SP4	Qingdao	DP10, DP11
	SP5	Tianjin	DP5
	SP6	Dailian	DP9
	SP7	Yingkou	DP9
	SP8	Lianyungang	DP8
	SP9	Rizhao	DP10, DP11
	SP10	Zhanjiang	DP3
	SP11	Guangzhou	DP3
Cross-border train station (CBT)	CBT1	Urumqi	DP12, DP2, DP13
	CBT2	Beijing	DP5, DP10
	CBT3	Nanning	DP3, DP7
	CBT4	Kashgar	DP13, DP12
	CBT5	Kunming	DP1
	CBT6	Lianyungang	DP6, DP8
	CBT7	Shenyang	DP9
	CBT8	Chongqing	DP1, DP6
	CBT9	Jinhua	DP8
	CBT10	Xiangtan	DP7, DP6
	CBT11	Chengdu	DP1, DP6
	CBT12	Suzhou	DP8

6.5 Robustness check

In this section, the author includes two tests to check the robustness of the proposed ARMEC-GCS method’s performance.

6.5.1.1 Test 1: Comparing the ARMEC-GCS and GCS-only approach

In this test, the solution for dry port locations and their assigned coverage areas is derived based on the GCS-only approach, without using the ARMEC model to mine association rules between international demand patterns and supplying locations. The results can be seen in Figure 6.7 and in Table 6.7.

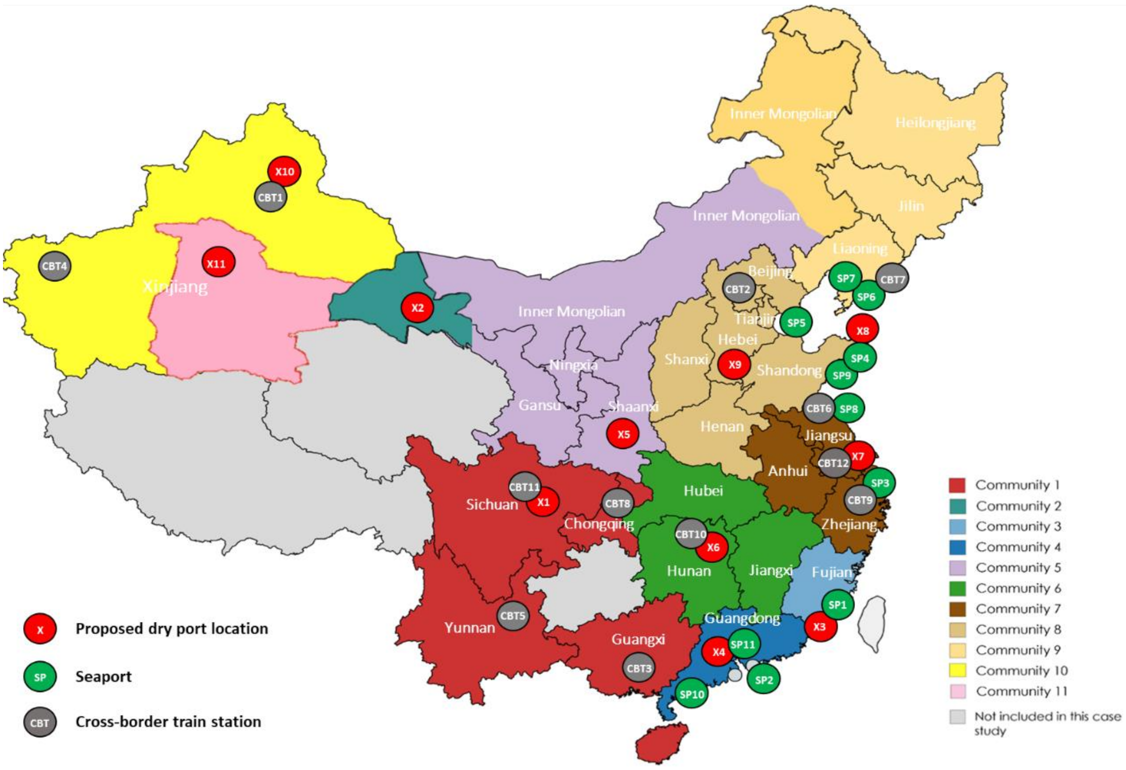


Figure 6.7 - Dry port locations based on the results of the GCS-only approach.

(Node codes are demonstrated in Tables 6.6 and 6.7)

Table 6.7 - Dry port locations with assigned communities using the GCS-only approach

Dry port code	Hub dry port location	Community size (Number of nodes)
X1	Chengdu	51
X2	Jiayuguan	2
X3	Xiamen	14
X4	Guangzhou	19
X5	Xi'an	39
X6	Xiangtan	32
X7	Suzhou	33
X8	Yantai	48
X9	Xingtai	57
X10	Urumqi	13
X11	Korla	1

As compared to the ARMEC-GCS, the GCS-only algorithm fails to suggest some key dry port locations, such as Beijing and Shenyang. As mentioned above, these two nodes are key nodes of the CMREC and the MSR. Without these dry port nodes, the wide inland areas of the Inner Mongolian, northern, and Metropolitan areas of China would easily fall into the disorder of logistics operations as the increasing volume of hinterland cargo from and to seaports will lead directly to severe traffic congestion, longer shipping times, and shortages of capacity at the seaports. The whole global shipping service would also suffer.

6.5.1.2 Test 2: Sensitivity analysis

In the ARMEC-GCS approach, two main parameters have an impact on the result: the minimum support and minimum confidence thresholds in the ARM model. Thus, the author conducts the sensitivity analysis by examining how the results would change when the values of these two parameters changed. The results can be seen in Table 6.8.

Table 6.8 - Sensitivity analysis result

Case	Support	Confidence	Number of rules	Number of cities
Case 1	0.000030	0.4	3110	80
Case 2	0.000030	0.7	1908	71
Case 3	0.000030	1	1837	71
Case 4	0.000100	0.4	737	53
Case 5	0.000100	0.7	353	44
Case 6	0.000100	1	186	37
Case 7	0.001000	0.4	51	16
Case 8	0.001000	0.7	13	10
Case 9	0.001000	1	5	5

As shown in Table 6.8, the number of association rules and number of cities ranked in the ARMEC model are quite sensitive to different settings of minimum support and minimum confidence thresholds. As the GCS model takes the ARMEC output as its input, the optimal locations and service areas of dry ports are also likely to change. In particular, when increasing the support and confidence thresholds, the number of rules drops significantly, which means less computation resources required. However, the number of cities also considerably reduces, which indicates information lost. Having a closer look at the dry port locations, the author notices that Beijing and Shenyang only appear in the ARMEC results in case 1 (Table 8), which supports thresholds set at the lowest possible value. This is because the Alibaba dataset is quite large and sparse, which is common in real-world data. Hence, the minimum support threshold needs to be as low as possible to avoid losing important information. However, doing that results in a large number of association rules, which require high computational costs. Thus, it is essential to develop a scalable visualization model that is capable of analysing the big set of association rules effectively. For that, the use of an EC-based complex network, as proposed in the ARMEC-GCS approach, is an effective way to enhance the scalability of the whole method.

6.6 Summary

On the basis of data mining and complex network analysis, this research proposes a two-stage ARMEC-GCS approach to optimise the location and service area of dry ports in a large-scale inland transportation system. In the first stage, the author uses ARM to extract, from a large transaction database, a set of association rules between international demand patterns and supplying locations. These association rules are then visualized as a complex network in which each supplying location is evaluated, with the EC score indicating its importance, weighted from international customers' point of view. In the second stage, the author adapts the classical FN algorithm from modularity-based community structure theory to propose the GCS algorithm, which optimises hub locations of dry ports and their coverage areas, based on inland regions' factors from the microeconomic (i.e., the EC score rankings), macroeconomic (i.e., foreign trade economics), and geographic (i.e., spatial distance) perspectives.

For model validation, the ARMEC-GCS approach is applied to optimise the dry port-based hinterland transportation network in Mainland China in the BRI context. In order to obtain the importance ranking of inland Chinese cities from the international demand perspective, the author constructs a real-world database recording 25,643 international Alibaba transactions in the BRI area. As a result, 309 Chinese cities from 24 inland provinces are grouped into 13 communities in which each dry port location suggested has strong hub functions in its feeder. Since the solution closely reflects several key BRI projects in both the SREB and MSR, the proposed method is validated as a reliable, scalable tool with which to seek optimal solutions for large-scale dry port location and allocation problems in practice.

There are some limitations in this research that should be investigated in future research. Firstly, the solution is quite sensitive to different settings of the minimum support and minimum confidence threshold in the ARM model. Hence, future research can improve the model reliability by feeding the optimization component into the ARM model to find optimal values for these parameters. Secondly, the proposed GCS algorithm adopts the hard network divisions for non-overlapping communities, meaning that an inland region can only belong to one community. It would be worthwhile in future studies to investigate dry port networks with overlapping communities, which are also very common in reality.

Chapter 7 Conclusion

This chapter concludes the thesis by providing a summary of the main findings, highlighting the theoretical and practical contributions of the study, as well as discussing the limitations and future research directions.

7.1 Summary of findings

This thesis aims to explore new applications of BDA to support the data-driven decision making in SCM. To do so, the thesis focuses on four research objectives and the summary of them are described as below.

(1) Conduct a literature review on the applications of BDA in SCM.

The rapid growing interest from both academics and practitioners towards the application of BDA in SCM has urged the need of review up-to-date research development in order to develop new agenda. This review responds to this call by proposing a novel classification framework that provides a full picture of current literature on where and how BDA has been applied within the SCM context. The classification framework is structured based on the content analysis method of Mayring (2008), addressing four research questions on (1) what areas in SCM that BDA is being applied, (2) what level of analytics is BDA used in these application areas, (3) what types of BDA models are used, and finally (4) what BDA techniques are employed to develop these models. The discussion tackling these four questions reveals a number of research gaps which leads to future research directions.

(2) Develop a comprehensible data-driven sales prediction model.

Remanufacturing has received increasing attention from researchers over the last decade. While many associated operational issues have been extensively studied, research into the prediction customer demand for, and the market development of, remanufactured products is still in its infancy. The majority of the existing research into remanufactured product demand is largely based on conventional statistical models that fail to capture the non-linear behaviour of customer demand and market factors in real-world business environments, in particular e-marketplaces. Therefore, this study aims to develop a comprehensible data-mining prediction approach, in order to achieve two objectives: (1) to provide a highly accurate and robust demand prediction model of remanufactured products; and (2) to shed light on the non-linear effect of online market factors as predictors of customer demand. Based on the real-world Amazon dataset, the results suggest that predicting remanufactured product demand is a complex, non-linear problem, and that, by using advanced machine-learning techniques, the proposed approach can predict the product demand with high accuracy. In terms of practical implications, the importance of market factors is ranked according to their predictive powers of demand, while their effects on demand are analysed through their partial dependence plots. Several insights for management are revealed by a thorough comparison of the sales impact of these market factors on remanufactured and new products.

(3) Develop a data-driven prescriptive optimisation model of promotional pricing.

To extract the most business value out of BDA, practitioners often perform predictive analytics in conjunction with prescriptive analytics to optimise the proactive decision making ahead of time. Hence, this study develops a two-stage, prescriptive promotional price optimisation empowered based on the data-driven sales forecasting. Particularly, the forecasted values of the product sales are first generated by using machine learning-based predictive analytics approach. Then, the price-expected sales data are inputted in the price optimisation model of which the objective is to determine the optimal promotional price of each product in order to maximise the total sales under some business constraints. The proposed approach is validated using the real-world database of 81,380 sales transactions. The result shows that the proposed prescriptive approach is able to provide the optimal promotional pricing solution that could increase both sales and revenues.

(4) Develop a data-driven prescriptive optimisation model of large-scale dry port network design.

The research proposes a two-stage approach that combines data mining and complex network theory to optimise the locations and service areas of dry ports in a large-scale inland transportation system. In the first stage, candidate locations of dry ports are weighted based on their eigenvector centrality in the complex network of association rules mined from a large amount of international transaction data. In the second phrase, dry port locations and their service areas are optimised using the gravity-based community structure. The method is validated in a real case study optimizing a large-scale dry port network in Mainland China in the context of the Belt and Road Initiatives (BRI). A real-world database of Alibaba transactions between China and other BRI countries is collected for association rule mining to represent international purchasing patterns. The result shows that the proposed model is able to provide realistic and applicable solutions for dry port developments, since it accurately pinpoints key locations in the real BRI development plans.

7.2 Contributions

The contribution of this thesis is multifaceted.

From theoretical perspectives:

- The literature review in this thesis is one of the earliest reviews in BDA-based SCM literature. It serves as a good starting point for researchers to build up the foundation of BDA and look for the potential research opportunities in the area.
- The three data-driven models proposed in this thesis all provide new ways to tackle the established problems in SCM (i.e, demand forecasting, price optimisation, logistics network design) with more industry-related findings than the conventional parametric approach.
- The methodological frameworks of these models are generalisable, adaptable, and extendable to apply for the other research area.

In terms of practical contributions:

- The literature review helps increase managers' understanding of what and how BDA can be used to extract business values in SCM areas, thereby increasing the practical use of BDA.
- The thesis facilitates the data-driven decision making across industries.

- The comprehensible predictive approach of remanufactured product demand can provide guidelines for managers to develop effective online marketing and selling strategies that takes into account the market behaviour of remanufactured products, thereby increasing the viability and profitability of remanufacturing/CLSCs.
- The prescriptive price optimisation is particularly beneficial to store managers who have to make a large number of daily promotional pricing decisions on what types of products put on sales and at what prices.
- The large-scale optimisation of Chinese dry port locations under the BRI initiative brings many benefits not only from international trading and transportation management perspectives but also from the political perspectives. Indeed, the optimal solution is realistic and applicable, as compared to the actual BRI development plan of dry ports in Mainland China.

From political perspectives, the solution for large-scale optimisation of Chinese dry port locations is derived from multiple decision-making perspectives (i.e., macroeconomic, microeconomic, and geographical), which in turn increases the possibility of its acceptance by various groups of stakeholders and of obtaining funds from the BRI investment.

7.3 Limitations and future directions

This thesis adopts data-driven decision making approach, allowing data to speak itself at the minimum impact of simplifying assumptions. However, it does not necessarily mean that the data-driven, machine learning-based aproach always outperforms the traditional model-driven, statistical approach. For example, as can be seen in Chapter 3 when we compare the demand prediction accuracy of Linear Regression (LR) with different machine learning models, LR outpeforms CART, M5 and ANN in many cases. Indeed, one of the main limitation of the data-driven methodology is that the obtained result is largely dependent on the quality of data collection and data preparation. The performance of machine learning models can also vary greatly acording to the nature of the data itself. Therefore, it is highly recommended that in order to find the best solution, practicioners should make comparisions between different methods and provide robustness checks rather than blindly apply machine learning approach in every case.

Furthermore, there are some limitations in this thesis which should be further examined in future research. First, one of the limitations in the literature review in Chapter 2 is that the categorisation in classification framework remaining interpretative, which could lead to the concern on subjective bias. This is also one of the well-established issues of the content analysis method despite a number of validation being done. However, the author has minimised the subjective bias by asking his supervisors to independently conduct the article selection process, and then comparing the reference list before generating the final decision.

Secondly, the main limitation in the predictive analytics of remanufactured product sales is that it excluded the time-series analysis of product sales. Therefore, future research can improve the prediction accuracy by integrating the different temporal data points of sales as input variables.

Thirdly, the main limitation in the prescriptive promotional price optimisation is about the problem size it can handle at its second stage of optimisation. The optimisation is formulated based on MILP approach to acquire exact solutions. However, the practical value of MILP-based models is still limited, especially when dealing with a real-world, complex system which inherently involves millions of integer decision variables. It can easily make the MILP model computationally intractable. Hence, future research can incorporate the sales predictive model at the first stage with the heuristic, approximation-based algorithm at the second stage to be able to handle the large-scale optimisation problem.

Finally, the data-driven prescriptive optimisation of dry port location also has some limitations. The obtained solution is quite sensitive to the change in the predefined minimum support and minimum confidence threshold in the ARM model at the first stage. Therefore, future research can improve the model by injecting the optimization component into the ARM model to determine optimal values for these parameters. Also, at the second stage, the community structure-based optimisation assumes the hard network divisions for non-overlapping communities, meaning that an inland region can only belong to one community. Hence, it would be worthwhile for future research to develop the optimisation model that allows overlapping communities which are also a very common phenomenon in reality.

References

Abbasi, M. and Pishvaee, M.S. (2018), "A two-stage GIS-based optimization model for the dry port location problem : A case study of Iran", *Journal of Industrial and Systems Engineering*.

Abbey, J.D., Blackburn, J.D. and Guide, V.D.R. (2015), "Optimal pricing for new and remanufactured products", *Journal of Operations Management*, No longer published by Elsevier, Vol. 36, pp. 130–146.

Abbey, J.D., Meloy, M.G., Guide, V.D.R. and Atalay, S. (2015), "Remanufactured Products in Closed-Loop Supply Chains for Consumer Goods", *Production and Operations Management*, John Wiley & Sons, Ltd (10.1111), Vol. 24 No. 3, pp. 488–503.

Achenbach, A. and Spinler, S. (2018), "Prescriptive analytics in airline operations: Arrival time prediction and cost index optimization for short-haul flights", *Operations Research Perspectives*, Elsevier, Vol. 5 No. July, pp. 265–279.

Addo-Tenkorang, R. and Helo, P.T. (2016), "Big Data Applications in Operations/Supply-Chain Management: A Literature Review", *Computers & Industrial Engineering*, Elsevier Ltd, p. .

Ahiaga-Dagbui, D.D. and Smith, S.D. (2014), "Dealing with construction cost overruns using data mining", *Construction Management and Economics*, Vol. 32 No. 7–8, pp. 682–694.

Akerlof, G.A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics*, Narnia, Vol. 84 No. 3, pp. 488–500.

Alaei, A.R., Becken, S. and Stantic, B. (2017), "Sentiment Analysis in Tourism: Capitalizing on Big Data", *Journal of Travel Research*, SAGE PublicationsSage CA: Los Angeles, CA, Vol. 58 No. 2, pp. 175–191.

Alic, A.S., Almeida, J., Aloisio, G., Andrade, N., Antunes, N., Ardagna, D., Badia, R.M., et al. (2019), "BIGSEA: A Big Data analytics platform for public transportation information”, *Future Generation Computer Systems*, North-Holland, Vol. 96, pp. 243–269.

Alqahtani, A.Y. and Gupta, S.M. (2017), “Evaluating two-dimensional warranty policies for remanufactured products”, *Journal of Remanufacturing*, Springer Netherlands, Vol. 7 No. 1, pp. 19–47.

Alyahya, S., Wang, Q. and Bennett, N. (2016), “Application and integration of an RFID-enabled warehousing management system – a feasibility study”, *Journal of Industrial Information Integration*, Elsevier Inc., Vol. 4, pp. 15–25.

Anderson, J.E. and van Wincoop, E. (2003), “Gravity with gravitas: a solution to the border puzzle”, *American Economic Review*, Vol. 93, pp. 170–192.

Archak, N., Ghose, A. and Ipeirotis, P.G. (2011), “Deriving the Pricing Power of Product Features by Mining Consumer Reviews”, *Management Science*, INFORMS , Vol. 57 No. 8, pp. 1485–1509.

Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M. a. S. and Buyya, R. (2015), “Big Data computing and clouds: Trends and future directions”, *Journal of Parallel and Distributed Computing*, Elsevier Inc., Vol. 79–80, pp. 3–15.

Atasu, A., Guide, V.D.R. and Van Wassenhove, L.N. (2008), “Product Reuse Economics in Closed-Loop Supply Chain Research”, *Production and Operations Management*, John Wiley & Sons, Ltd (10.1111), Vol. 17 No. 5, pp. 483–496.

Atasu, A., Guide, V.D.R. and Van Wassenhove, L.N. (2010), “So What If Remanufacturing Cannibalizes My New Product Sales?”, *California Management Review*, SAGE PublicationsSage CA: Los Angeles, CA, Vol. 52 No. 2, pp. 56–76.

Babiceanu, R.F. and Seker, R. (2016), “Big Data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook”, *Computers in Industry*, Elsevier B.V., Vol. 81, pp. 128–137.

Bae, J.K. and Kim, J. (2011), “Product development with data mining techniques: A case on design of digital camera", *Expert Systems with Applications*, Elsevier Ltd, Vol. 38 No. 8, pp. 9274–9280.

Ballestín, F., Pérez, Á., Lino, P., Quintanilla, S. and Valls, V. (2013), "Static and dynamic policies with RFID for the scheduling of retrieval and storage warehouse operations", *Computers & Industrial Engineering*, Vol. 66 No. 4, pp. 696–709.

Belaud, J.-P., Prioux, N., Vialle, C. and Sablayrolles, C. (2019), "Big data for agri-food 4.0: Application to sustainability management for by-products supply chain", *Computers in Industry*, Elsevier, Vol. 111, pp. 41–50.

Belt & Road News. (2019), "New Freight Train Route links Yantai, Duisburg - Belt &amp; Road News", *Belt & Road News*, available at: https://www.beltandroad.news/2019/07/28/new-freight-train-route-links-yantai-duisburg/ (accessed 14 September 2019).

Belyi, E., Giabbanelli, P.J., Patel, I., Balabhadrapathruni, N.H., Abdallah, A. Ben, Hameed, W. and Mago, V.K. (2016), "Combining association rule mining and network analysis for pharmacosurveillance", *Journal of Supercomputing*, Vol. 72 No. 5, pp. 2014–2034.

Berengueres, J. and Efimov, D. (2014), "Airline new customer tier level forecasting for real-time resource allocation of a miles program", *Journal Of Big Data*, Vol. 1 No. 1, p. 3.

De Bock, K.W. (2017), "The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles", *Expert Systems with Applications*, Pergamon, Vol. 90, pp. 23–39.

Bogomolova, S., Szabo, M. and Kennedy, R. (2017), "Retailers' and manufacturers' price-promotion decisions: Intuitive or evidence-based?", *Journal of Business Research*, Elsevier Inc., Vol. 76, pp. 189–200.

Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. (2017), "Explaining machine learning models in sales predictions", *Expert Systems with Applications*, Pergamon, Vol. 71, pp. 416–428.

Boulding, W. and Kirmani, A. (1993), “A Consumer-Side Experimental Examination of Signaling Theory: Do Consumers Perceive Warranties as Signals of Quality?”, *Journal of Consumer Research*, Oxford University Press, Vol. 20, pp. 111–123.

Breiman, L. (2001), “Random Forests”, *Machine Learning*, Kluwer Academic Publishers, Vol. 45 No. 1, pp. 5–32.

Bryman, A. and Bell, E. (2015), *Business Research Methods*, 4th ed., OUP Oxford.

De Caigny, A., Coussement, K. and De Bock, K.W. (2018), “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees”, *European Journal of Operational Research*, North-Holland, Vol. 269 No. 2, pp. 760–772.

Çalı, S. and Balaman, Ş.Y. (2019), “Improved decisions for marketing, supply and purchasing: Mining big data through an integration of sentiment analysis and intuitionistic fuzzy multi criteria assessment”, *Computers & Industrial Engineering*, Pergamon, Vol. 129, pp. 315–332.

Campbell, J.F. and O’Kelly, M.E. (2012), “Twenty-Five Years of Hub Location Research”, *Transportation Science*, Vol. 46 No. 2, pp. 153–169.

Canh, L. and Notteboom, T. (2016), “A Multi-Criteria Approach to Dry Port Location in Developing Economies with Application to Vietnam”, *The Asian Journal of Shipping and Logistics*.

Caro, F. and Gallien, J. (2012), “Clearance Pricing Optimization for a Fast-Fashion Retailer”, *Operations Research*, Vol. 60 No. 6, pp. 1404–1422.

Cavalcante, I.M., Frazzon, E.M., Forcellini, F.A. and Ivanov, D. (2019), “A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing”, *International Journal of Information Management*, Pergamon, Vol. 49, pp. 86–97.

Chang, B.R., Tsai, H.-F. and Liao, P.-H. (2018), “Applying intelligent data traffic adaptation to high-performance multiple big data analytics platforms”, *Computers & Electrical Engineering*, Pergamon, Vol. 70, pp. 998–1018.

Chang, Z., Notteboom, T. and Lu, J. (2015), “A two-phase model for dry port location with an application to the port of Dalian in China”, *Transportation Planning and Technology*, Vol. 38 No. 4, pp. 442–464.

Chase, C.W. (2016), *Next Generation Demand Management: People, Process, Analytics, and Technology*, Wiley Online Library.

Chen, H., Lam, J.S.L. and Liu, N. (2018), “Strategic investment in enhancing port–hinterland container transportation network resilience: A network game theory approach”, *Transportation Research Part B: Methodological*, Vol. 111, pp. 83–112.

Chevalier, J.A. and Mayzlin, D. (2006), “The Effect of Word of Mouth on Sales: Online Book Reviews”, *Journal of Marketing Research*, SAGE PublicationsSage CA: Los Angeles, CA, Vol. 43 No. 3, pp. 345–354.

Chiang, D.M.-H., Lin, C.-P. and Chen, M.-C. (2011), “The adaptive approach for storage assignment by mining data of warehouse management system for distribution centres”, *Enterprise Information Systems*, Vol. 5 No. 2, pp. 219–234.

Chiang, D.M.-H., Lin, C.-P. and Chen, M.-C. (2014), “Data mining based storage assignment heuristics for travel distance reduction”, *Expert Systems*, Vol. 31 No. 1, pp. 81–90.

Chien, C.-F., Diaz, A.C. and Lan, Y.-B. (2014), “A data mining approach for analyzing semiconductor MES and FDC data to enhance overall usage effectiveness (OUE)”, *International Journal of Computational Intelligence Systems*, Vol. 7 No. sup2, pp. 52–65.

China Daily. (2015), “Guangdong sees big role in ‘One Belt, One Road’ - China - Chinadaily.com.cn”, *China Daily*, available at: http://www.chinadaily.com.cn/regional/2015-06/05/content_20930718.htm (accessed 15 September 2019).

China Daily. (2017), “Yantai awarded as one of most dynamic Belt and Road cities”, *China Daily,* available at: http://www.chinadaily.com.cn/m/shandong/yantai/2017-01/16/content_27962935.htm (accessed 14 September 2019).

China Daily. (2018), “Guangdong – A key hub on the Maritime Silk Road - Opinion - Chinadaily.com.cn”, *China Daily,* available at: http://www.chinadaily.com.cn/a/201809/25/WS5ba9e159a310c4cc775e7fbc.html (accessed 15 September 2019).

Choi, Y., Lee, H. and Irani, Z. (2016), “Big data-driven fuzzy cognitive map for prioritising IT service procurement in the public sector”, *Annals of Operations Research,* Springer US, pp. 1–30.

Chong, A.Y.L., Li, B., Ngai, E.W.T., Ch’ng, E. and Lee, F. (2016), “Predicting online product sales via online reviews, sentiments, and promotion strategies”, *International Journal of Operations & Production Management,* Emerald Group Publishing Limited , Vol. 36 No. 4, pp. 358–383.

Chuang, Y.F., Chia, S.H. and Wong, J.Y. (2014), “Enhancing order-picking efficiency through data mining and assignment approaches”, *WSEAS Transactions on Business and Economics,* Vol. 11 No. 1, pp. 52–64.

Cialdini, R.B. (2009), *Influence: Science and Practice,* 5th ed., Pearson.

Clauset, A., Newman, M.E.J. and Moore, C. (2004), “Finding community structure in very large networks”, *Physical Review E,* Vol. 70 No. 6, p. 6.

Corne, D., Dhaenens, C. and Jourdan, L. (2012), “Synergies between operations research and data mining: The emerging use of multi-objective approaches”, *European Journal of Operational Research,* Vol. 221 No. 3, pp. 469–479.

Costa, A. (2015), “MILP formulations for the modularity density maximization problem”, *European Journal of Operational Research,* Elsevier Ltd., Vol. 245 No. 1, pp. 14–21.

Crainic, T.G., Dell’Olmo, P., Ricciardi, N. and Sgalambro, A. (2015), “Modeling dry-port-based freight distribution planning", *Transportation Research Part C: Emerging Technologies,* Vol. 55, pp. 518–534.

Cui, G., Lui, H.-K. and Guo, X. (2012), "The Effect of Online Consumer Reviews on New Product Sales", *International Journal of Electronic Commerce,* Routledge , Vol. 17 No. 1, pp. 39–58.

Cui, J., Liu, F., Hu, J., Janssens, D., Wets, G. and Cools, M. (2016), "Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast growing Chinese city of Harbin", *Neurocomputing,* Elsevier, Vol. 181, pp. 4–18.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J. (2007), "Random forests for classification in ecology", *Ecology,* John Wiley & Sons, Ltd, Vol. 88 No. 11, pp. 2783–2792.

Dai, Q., Zhong, R., Huang, G.Q., Qu, T., Zhang, T. and Luo, T.Y. (2012), "Radio frequency identification-enabled real-time manufacturing execution system: a case study in an automotive part manufacturer", *International Journal of Computer Integrated Manufacturing,* Vol. 25 No. 1, pp. 51–65.

Dekkers, R. (2011), "Impact of strategic decision making for outsourcing on managing manufacturing", *International Journal of Operations & Production Management,* Emerald Group Publishing Limited, Vol. 31 No. 9, pp. 935–965.

Delen, D. and Demirkan, H. (2013), "Data, information and analytics as services", *Decision Support Systems,* Elsevier B.V., Vol. 55 No. 1, pp. 359–363.

Delen, D., Erraguntla, M., Mayer, R.J. and Wu, C.N. (2011), "Better management of blood supply-chain with GIS-based analytics", *Annals of Operations Research,* Vol. 185 No. 1, pp. 181–193.

Dev, N.K., Shankar, R., Gupta, R. and Dong, J. (2019), "Multi-criteria evaluation of real-time key performance indicators of supply chain with consideration of big data architecture", *Computers & Industrial Engineering*, Pergamon, Vol. 128, pp. 1076–1087.

Dey, S., Gupta, N., Pathak, S., Kela, D.H. and Datta, S. (2017), "Data-Driven Design Optimization for Industrial Products", *Optimization in Industry*, Springer International Publishing, pp. 253–267.

DHL. (2016), *"Belt and Road": What You Need to Know*, available at: https://www.logistics.dhl/content/dam/dhl/global/dhl-global-forwarding/documents/pdf/dhl-glo-dgf-belt-and-road.pdf (accessed 14 September 2019).

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006), "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, BioMed Central, Vol. 7 No. 1, p. 3.

Do, N. (2014), "Application of OLAP to a PDM database for interactive performance evaluation of in-progress product development", *Computers in Industry*, Elsevier B.V., Vol. 65 No. 4, pp. 636–645.

Dobre, C. and Xhafa, F. (2014), "Intelligent services for Big Data science", *Future Generation Computer Systems*, Elsevier B.V., Vol. 37, pp. 267–281.

Duan, L. and Xiong, Y. (2015), "Big data analytics and business analytics", *Journal of Management Analytics*, Taylor & Francis, Vol. 2 No. 1, pp. 1–21.

Dutta, D. and Bose, I. (2015), "Managing a Big Data project: The case of Ramco Cements Limited", *International Journal of Production Economics*, Elsevier, Vol. 165, pp. 293–306.

Ehmke, J.F., Campbell, A.M. and Thomas, B.W. (2016), "Data-driven approaches for emissions-minimized paths in urban areas", *Computers and Operations Research*, Elsevier, Vol. 67, pp. 34–47.

Emani, C.K., Cullot, N. and Nicolle, C. (2015), "Understandable Big Data: A survey", *Computer Science Review*, Elsevier Inc., Vol. 17, pp. 70–81.

en.people.cn. (2018), "The Belt and Road gives boost to Ningbo-Zhoushan port - People's Daily Online", *En.People.Cn*, available at: http://en.people.cn/n3/2018/0814/c90000-9490588.html (accessed 14 September 2019).

Erl, T., Khattak, W. and Buhler, P. (2016), *Big Data Fundamentals*, PRENTICE HALL.

Fang, X. and Zhan, J. (2015), "Sentiment analysis using product review data", *Journal of Big Data*, Journal of Big Data, Vol. 2 No. 1, p. 5.

Feng, X., Zhang, Y., Li, Y. and Wang, W. (2013), "A location-allocation model for seaport-dry port system optimization", *Discrete Dynamics in Nature and Society*, Vol. 2013, available at:https://doi.org/10.1155/2013/309585.

Ferreira, K.J., Lee, B.H.A. and Simchi-Levi, D. (2016), "Analytics for an Online Retailer: Demand Forecasting and Price Optimization", *Manufacturing & Service Operations Management*, Vol. 18 No. 1, pp. 69–88.

Fischetti, M., Ljubić, I. and Sinnl, M. (2017), "Redesigning Benders Decomposition for Large-Scale Facility Location", *Management Science*, Vol. 63 No. 7, pp. 2146–2162.

Frota Neto, J.Q., Bloemhof, J. and Corbett, C. (2016), "Market prices of remanufactured, used and new items: Evidence from eBay", *International Journal of Production Economics*, Elsevier, Vol. 171, pp. 371–380.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, Elsevier Ltd, Vol. 35 No. 2, pp. 137–144.

Ghanbari, R., Jalili, M. and Yu, X. (2018), "Correlation of cascade failures and centrality measures in complex networks", *Future Generation Computer Systems*, Vol. 83, pp. 390–400.

Ghedini Ralha, C. and Sarmento Silva, C.V. (2012), "A multi-agent data mining system for cartel detection in Brazilian government procurement", *Expert Systems with Applications*, Vol. 39 No. 14, pp. 11642–11656.

Giannakis, M. and Louis, M. (2016), "A multi-agent based system with big data processing for enhanced supply chain agility", *Journal of Enterprise Information Management*, Vol. 29 No. 5, pp. 706–727.

Gohar, M., Muzammal, M. and Ur Rahman, A. (2018), "SMART TSS: Defining transportation system behavior using big data analytics in smart cities", *Sustainable Cities and Society*, Elsevier, Vol. 41, pp. 114–119.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015), "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation", *Journal of Computational and Graphical Statistics*, Taylor & Francis, Vol. 24 No. 1, pp. 44–65.

Govindan, K., Soleimani, H. and Kannan, D. (2015), "Reverse logistics and closed-loop supply chain: A comprehensive review to explore the future", *European Journal of Operational Research*, Vol. 240 No. 3, pp. 603–626.

Grömping, U. (2015), "Variable importance in regression models", *Wiley Interdisciplinary Reviews: Computational Statistics*, John Wiley & Sons, Ltd, Vol. 7 No. 2, pp. 137–152.

Guba, E.G. (1990), *The Paradigm Dialog*, Sage Publications.

Guide Jr., V.D.R. and Li, J. (2010), "The Potential for Cannibalization of New Products Sales by Remanufactured Products*", *Decision Sciences*, John Wiley & Sons, Ltd (10.1111), Vol. 41 No. 3, pp. 547–572.

Guo, S.Y., Ding, L.Y., Luo, H.B. and Jiang, X.Y. (2016a), "A Big-Data-based platform of workers' behavior: Observations from the field", *Accident Analysis & Prevention*, Elsevier Ltd, Vol. 93, pp. 299–309.

Guo, S.Y., Ding, L.Y., Luo, H.B. and Jiang, X.Y. (2016b), "A Big-Data-based platform of workers' behavior: Observations from the field", *Accident Analysis and Prevention*, Elsevier Ltd, Vol. 93, pp. 299–309.

Hamzaoui-Essoussi, L. and Linton, J.D. (2014), "Offering branded remanufactured/recycled products: at what price?”, *Journal of Remanufacturing*, Springer Netherlands, Vol. 4 No. 1, p. 9.

Hamzaoui Essoussi, L. and Linton, J.D. (2010), “New or recycled products: how much are consumers willing to pay?”, *Journal of Consumer Marketing*, Emerald Group Publishing Limited, Vol. 27 No. 5, pp. 458–468.

Harsha, P., Subramanian, S. and Ettl, M. (2019), “A Practical Price Optimization Approach for Omni-channel Retailing”, *INFORMS Journal on Optimization*, pp. 1–45.

Hazen, B.T., Skipper, J.B., Boone, C.A. and Hill, R.R. (2016), “Back in business: operations research in support of big data analytics for operations and supply chain management”, *Annals of Operations Research*, Springer US, pp. 1–11.

He, W., Wu, H., Yan, G., Akula, V. and Shen, J. (2015), “A novel social media competitive analytics framework with sentiment benchmarks”, *Information & Management*, North-Holland, Vol. 52 No. 7, pp. 801–812.

van Heijst, D., Potharst, R. and van Wezel, M. (2008), “A support system for predicting eBay end prices”, *Decision Support Systems*, North-Holland, Vol. 44 No. 4, pp. 970–982.

Henttu, V. and Hilmola, O. (2011), “Financial and environmental impacts of hypothetical Finnish dry port structure”, *Research in Transportation Economics*, Vol. 33, pp. 35–41.

Hey, T., Tansley, S. and Tolle, K. (2009), *The Fourth Paradigm : Data-Intensive Scientific Discovery*, 1st ed., Microsoft Research.

HKTDC research. (2019), “Guangdong: Market Profile”, *HKTDC Research*, available at: http://china-trade-research.hktdc.com/business-news/article/Facts-and-Figures/Guangdong-Market-Profile/ff/en/1/1X000000/1X06BUOU.htm (accessed 15 September 2019).

Ho, S.Y. and Bodoff, D. (2014), “The Effects of Web Personalization on User Attitude and Behavior: An Integration of the Elaboration Likelihood Model and Consumer Search Theory”, *MIS Quarterly*, Vol. 38 No. 2, pp. 497–520.

Hofmann, E. (2015), “Big data and supply chain decisions: the impact of volume, variety and velocity properties on the bullwhip effect”, *International Journal of Production Research,* Vol. 7543 No. December 2015, pp. 1–19.

Hsu, C.-Y., Lin, S.-C. and Chien, C.-F. (2015), “A back-propagation neural network with a distributed lag model for semiconductor vendor-managed inventory”, *Journal of Industrial and Production Engineering*, Vol. 32 No. 3, pp. 149–161.

Hsu, C.-Y., Yang, C.-S., Yu, L.-C., Lin, C.-F., Yao, H.-H., Chen, D.-Y., Robert Lai, K., et al. (2015), “Development of a cloud-based service framework for energy conservation in a sustainable intelligent transportation system”, *International Journal of Production Economics,* Elsevier, Vol. 164, pp. 454–461.

Hu, N., Koh, N.S. and Reddy, S.K. (2014), “Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales”, *Decision Support Systems,* North-Holland, Vol. 57, pp. 42–53.

Huang, T., Lan, L., Fang, X., An, P., Min, J. and Wang, F. (2015), “Promises and Challenges of Big Data Computing in Health Sciences”, *Big Data Research,* Elsevier Inc., Vol. 2 No. 1, pp. 2–11.

Huang, T. and Van Mieghem, J.A. (2014), “Clickstream data and inventory management: Model and empirical analysis”, *Production and Operations Management,* Vol. 23 No. 3, pp. 333–347.

Huang, Y.-Y. and Hanfield, R.B. (2015), “Measuring the benefits of ERP on supply management maturity model: a ‘big data’ method”, *International Journal of Operations & Production Management,* Vol. 35 No. 1, pp. 2–25.

Huang, Y. (2016), “Understanding China’s Belt & Road Initiative: Motivation, framework and assessment”, *China Economic Review,* Vol. 40, pp. 314–321.

IBM. (2012), “What is Big Data?.”, *IBM Corporation*.

IDC. (2012), *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest*

Growth in the Far East.

Ijadi Maghsoodi, A., Kavian, A., Khalilzadeh, M. and Brauers, W.K.M. (2018), “CLUS-MCDA: A novel framework based on cluster analysis and multiple criteria decision theory in a supplier selection problem”, *Computers & Industrial Engineering*, Pergamon, Vol. 118, pp. 409–422.

Ishfaq, R. and Sox, C.R. (2011), “Hub location-allocation in intermodal logistic networks”, *European Journal of Operational Research*, Vol. 210 No. 2, pp. 213–230.

Ito, S. and Fujimaki, R. (2017), “Optimization Beyond Prediction: Prescriptive Price Optimization”, *KDD 2017 Applied Data Science*, pp. 1833–1841.

Jain, R., Singh, A.R., Yadav, H.C. and Mishra, P.K. (2014), “Using data mining synergies for evaluating criteria at pre-qualification stage of supplier selection”, *Journal of Intelligent Manufacturing*, Vol. 25 No. 1, pp. 165–175.

Jakowczyk, M., Frota Neto, J.Q., Gibson, A. and Van Wassenhove, L.N. (2017), “Understanding the market for remanufactured products: what can we learn from online trading and Web search sites?”, *International Journal of Production Research*, Taylor & Francis, Vol. 55 No. 12, pp. 3465–3479.

Jiao, Z., Ran, L., Zhang, Y., Li, Z. and Zhang, W. (2018), “Data-driven approaches to integrated closed-loop sustainable supply chain design under multi-uncertainties”, *Journal of Cleaner Production*, Elsevier Ltd, Vol. 185, pp. 105–127.

Jiménez-Parra, B., Rubio, S. and Vicente-Molina, M.-A. (2014), “Key drivers in the behavior of potential consumers of remanufactured products: a study on laptops in Spain”, *Journal of Cleaner Production*, Elsevier, Vol. 85, pp. 488–496.

Jin, J., Liu, Y., Ji, P. and Liu, H. (2016), “Understanding big consumer opinion data for market-driven product design”, *International Journal of Production Research*, Vol. 54 No. 10, pp. 3019–3041.

Jun, S.-P., Park, D.-H. and Yeom, J. (2014), “The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference", *Technological Forecasting and Social Change*, Elsevier Inc., Vol. 86, pp. 237–253.

Ka, B. (2011), "Application of fuzzy AHP and ELECTRE to China dry port location selection", *Asian Journal of Shipping and Logistics*, Vol. 27 No. 2, pp. 331–354.

Kakatkar, C. and Spann, M. (2019), "Marketing analytics using anonymized and fragmented tracking data", *International Journal of Research in Marketing*, North-Holland, Vol. 36 No. 1, pp. 117–136.

Kaluza, P., Kölzsch, A., Gastner, M.T. and Blasius, B. (2010), "The complex network of global cargo ship movements", *Journal of the Royal Society Interface*, pp. 1093–1103.

Kaur, H. and Singh, S.P. (2018), "Heuristic modeling for sustainable procurement and logistics in a supply chain using big data", *Computers & Operations Research*, Pergamon, Vol. 98, pp. 301–321.

Keeves, J.P. (1997), *Educational Research, Methodology and Measurement : An International Handbook*, Vol. 7, Pergamon, available at: https://books.emeraldinsight.com/page/detail/Educational-Research,-Methodology-and-Measurement/?k=9780080427102 (accessed 1 September 2019).

Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G. and Hooker, G. (2009), "Data-intensive Science: A New Paradigm for Biodiversity Studies", *BioScience*, Narnia, Vol. 59 No. 7, pp. 613–620.

Khor, K.S. and Hazen, B.T. (2017), "Remanufactured products purchase intentions and behaviour: Evidence from Malaysia", *International Journal of Production Research*, Taylor & Francis, Vol. 55 No. 8, pp. 2149–2162.

khosravi, S. and Akbari Jokar, M.R. (2017), "Facility and hub location model based on gravity rule", *Computers and Industrial Engineering*, Elsevier Ltd, Vol. 109, pp. 28–38.

Kitchin, R. (2014), "Big Data, new epistemologies and paradigm shifts", *Big Data & Society*, Vol. 1 No. 1, p. 205395171452848.

Kivunja, C. and Kuyini, A.B. (2017), “Understanding and Applying Research Paradigms in Educational Contexts”, *International Journal of Higher Education*, Vol. 6 No. 5, p. 26.

Kohavi, R. (1995), “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137–1145.

Komchornrit, K. (2017), “The Selection of Dry Port Location by a Hybrid CFA-MACBETH-PROMETHEE Method : A Case Study of Southern Thailand”, *The Asian Journal of Shipping and Logistics*, Vol. 33 No. 3, pp. 141–153.

Konings, J. (2018), *Trade Impacts of the Belt and Road Initiative*, available at: www.ing.com/THINK (accessed 29 March 2019).

KPMG China. (2018), “A New Xi’an in the New Era — KPMG Contributes to the Development of a New Reform Locomotive in Inland China - KPMG China”, *KPMG China*, 29 June, available at: https://home.kpmg/cn/en/home/news-media/press-releases/2018/06/a-new-xi-an-in-the-new-era.html (accessed 13 September 2019).

Krauss, C., Do, X.A. and Huck, N. (2017), “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&amp;P 500”, *European Journal of Operational Research*, North-Holland, Vol. 259 No. 2, pp. 689–702.

Krumeich, J., Werth, D. and Loos, P. (2015), “Prescriptive Control of Business Processes”, *Business & Information Systems Engineering*, Springer Fachmedien Wiesbaden, available at:https://doi.org/10.1007/s12599-015-0412-2.

Kumar, A.., Shankar, R.., Choudhary, A.. and Thakur, L.S.. (2016a), “A big data MapReduce framework for fault diagnosis in cloud-based manufacturing”, *International Journal of Production Research*, Vol. 7543 No. November, pp. 1–14.

Kumar, A., Shankar, R. and Aljohani, N.R. (2019), “A big data driven framework for demand-driven forecasting with effects of marketing-mix variables”, *Industrial Marketing Management*, Elsevier, available at:https://doi.org/10.1016/J.INDMARMAN.2019.05.003.

Kumar, A., Shankar, R., Choudhary, A. and Thakur, L.S. (2016b), "A big data MapReduce framework for fault diagnosis in cloud-based manufacturing", *International Journal of Production Research*, Vol. 54 No. 23, pp. 7060–7073.

Kunc, M. and O'Brien, F.A. (2018), "The role of business analytics in supporting strategy processes: Opportunities and limitations", *Journal of the Operational Research Society*, Taylor & Francis, Vol. 70 No. 6, pp. 974–985.

Kuntner, T. and Teichert, T. (2016), "The scope of price promotion research: An informetric study", *Journal of Business Research*, Elsevier Inc., Vol. 69 No. 8, pp. 2687–2696.

Kunz, T.P. and Crone, S.F.C. (2014), "Demand models for the static retail price optimization problem - A Revenue Management perspective", *SCOR*, pp. 101–125.

Kuo, R.J., Pai, C.M., Lin, R.H. and Chu, H.C. (2015), "The integration of association rule mining and artificial immune network for supplier selection and order quantity allocation", *Applied Mathematics and Computation*, Elsevier Inc., Vol. 250, pp. 958–972.

Kursa, M.B. and Rudnicki, W.R. (2010), "Feature Selection with the **Boruta** Package", *Journal of Statistical Software*, Vol. 36 No. 11, pp. 1–13.

De La Iglesia, B., Richards, G., Philpott, M.S. and Rayward-Smith, V.J. (2006), "The application and effectiveness of a multi-objective metaheuristic algorithm for partial classification", *European Journal of Operational Research*, Vol. 169 No. 3, pp. 898–917.

Lather, P. (1986), "Research as Praxis", *Harvard Educational Review*, Harvard Education Publishing Group , Vol. 56 No. 3, pp. 257–278.

Lättilä, L., Henttu, V. and Hilmola, O. (2015), "Hinterland operations of sea ports do matter : Dry port usage effects on transportation costs and CO 2 emissions", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 55 No. 2013, pp. 23–42.

Lee, C.K.H. (2016), “A GA-based optimisation model for big data analytics supporting anticipatory shipping in Retail 4.0”, *International Journal of Production Research*, Vol. 7543, pp. 1–13.

Lee, C.K.H., Choy, K.L., Ho, G.T.S. and Lin, C. (2016), “A cloud-based responsive replenishment system in a franchise business model using a fuzzy logic approach”, *Expert Systems*, Vol. 33 No. 1, pp. 14–29.

Lee, C.Y. and Song, D.P. (2017), “Ocean container transport in global supply chains: Overview and research opportunities”, *Transportation Research Part B: Methodological*, Elsevier Ltd, Vol. 95, pp. 442–474.

Lehman Brown International Accountants. (2017), *The Belt and Road Initiative, Lehman Brown International Accountants*, available at:https://doi.org/10.17265/2160-6579/2016.02.002.

Lei, N. and Moon, S.K. (2015), “A Decision Support System for market-driven product positioning and design”, *Decision Support Systems*, Elsevier B.V., Vol. 69, pp. 82–91.

Lepenioti, K., Bousdekis, A., Apostolou, D. and Mentzas, G. (2020), “Prescriptive analytics: Literature review and research challenges”, *International Journal of Information Management*, Elsevier, Vol. 50 No. October 2018, pp. 57–70.

Li, B., Ch’ng, E., Chong, A.Y.-L. and Bao, H. (2016a), “Predicting online e-marketplace sales performances: A big data approach”, *Computers & Industrial Engineering*, Elsevier Ltd, Vol. 101, pp. 565–571.

Li, B., Ch’ng, E., Chong, A.Y.-L. and Bao, H. (2016b), “Predicting online e-marketplace sales performances: A big data approach”, *Computers & Industrial Engineering*, Pergamon, Vol. 101, pp. 565–571.

Li, F., Shi, X. and Hu, H. (2011), “Location selection of dry port based on AP clustering: The case of SouthWest China”, *Journal of System and Management Sciences*, Vol. 2 No. 5, pp. 255–261.

Li, H., Fang, Y., Wang, Y., Lim, K.H. and Liang, L. (2015), "Are all signals equal? Investigating the differential effects of online signals on the sales performance of e-marketplace sellers", *Information Technology & People,* Emerald Group Publishing Limited , Vol. 28 No. 3, pp. 699–723.

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D. and Hampapur, A. (2014), "Improving rail network velocity: A machine learning approach to predictive maintenance", *Transportation Research Part C: Emerging Technologies,* Elsevier Ltd, Vol. 45, pp. 17–26.

Li, J., Moghaddam, M. and Nof, S.Y. (2016), "Dynamic storage assignment with product affinity and ABC classification???a case study", *International Journal of Advanced Manufacturing Technology,* The International Journal of Advanced Manufacturing Technology, Vol. 84 No. 9–12, pp. 2179–2194.

Li, L., Su, X., Wang, Y., Lin, Y., Li, Z. and Li, Y. (2014), "Robust causal dependence mining in big data network and its application to traffic flow predictions", *Transportation Research Part C: Emerging Technologies*, Vol. 58, pp. 292–307.

Li, X., Song, J. and Huang, B. (2016), "A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics", *International Journal of Advanced Manufacturing Technology,* Vol. 84 No. 1–4, pp. 119–131.

Ling Ho, C. and Wen Shih, H. (2014), "Applying Data Ming to Develop a Warning System of Procurement in Construction", *International Journal of Future Computer and Communication,* Vol. 3 No. 3, pp. 168–171.

Little, R.J.A. (1988), "A Test of Missing Completely at Random for Multivariate Data with Missing Values", *Journal of the American Statistical Association,* Taylor & Francis, Ltd.American Statistical Association, Vol. 83 No. 404, p. 1198.

Liu, J.-W. (2019), "Using big data database to construct new GFuzzy text mining and decision algorithm for targeting and classifying customers", *Computers & Industrial*

Engineering, Pergamon, Vol. 128, pp. 1088–1095.

Liu, X., Shin, H. and Burns, A.C. (2019), "Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing", *Journal of Business Research*, Elsevier, available at:https://doi.org/10.1016/J.JBUSRES.2019.04.042.

Liu, Y., Liu, G., Liu, Q. and Qin, Z. (2013), "Community Detection in Real Large Directed Weighted Networks", *International Journal of Digital Content Technology and Its Applications*, Vol. 7 No. 5, pp. 521–529.

Liu, Y.Q. and Wang, H. (2016), "Order allocation for service supply chain base on the customer best delivery time under the background of big data", *International Journal of Computer Science and Applications*, Vol. 13 No. 1, pp. 84–92.

Ma, J., Kwak, M. and Kim, H.M. (2014), "Demand trend mining for predictive life cycle design", *Journal of Cleaner Production*, Elsevier Ltd, Vol. 68, pp. 189–199.

Ma, S. and Fildes, R. (2017), "A retail store SKU promotions optimization model for category multi-period profit maximization", *European Journal of Operational Research*, Elsevier B.V., Vol. 260 No. 2, pp. 680–692.

De Maeyer, P. (2012), "Impact of online consumer reviews on sales and price strategies: a review and directions for future research", *Journal of Product & Brand Management*, Emerald Group Publishing Limited, Vol. 21 No. 2, pp. 132–139.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), *Big Data: The next Frontier for Innovation, Competition, and Productivity*, *McKinsey Global Institute*, available at:https://doi.org/10.1080/01443610903114527.

Marine-Roig, E. and Anton Clavé, S. (2015), "Tourism analytics with massive user-generated content: A case study of Barcelona", *Journal of Destination Marketing and Management*, Elsevier, Vol. 4 No. 3, pp. 162–172.

Martens, D., Baesens, B., Van Gestel, T. and Vanthienen, J. (2007), "Comprehensible credit scoring models using rule extraction from support vector machines", *European Journal of Operational Research*, North-Holland, Vol. 183 No. 3, pp. 1466–1476.

Masci, C., Johnes, G. and Agasisti, T. (2018), "Student and school performance across countries: A machine learning approach", *European Journal of Operational Research*, North-Holland, Vol. 269 No. 3, pp. 1072–1085.

Mayring, P. (2008), *Qualitative Inhaltsanalyse (Qualitative Content Analysis)*, 10th ed., Weinheim, Beltz.

Mazhar, M.I., Kara, S. and Kaebernick, H. (2007), "Remaining life estimation of used components in consumer products: Life cycle data analysis by Weibull and artificial neural networks", *Journal of Operations Management*, No longer published by Elsevier, Vol. 25 No. 6, pp. 1184–1193.

Mehmood, R., Meriton, R., Graham, G., Hennelly, P., Kumar, M., Mehmood, R., Meriton, R., et al. (2017), "Exploring the influence of big data on city transport operations : a Markovian approach", *Journal of Operations & Production Management*, Vol. 37 No. 1, pp. 75–104.

Meziane, F. and Proudlove, N. (2000), "Intelligent systems in manufacturing : current developments and future prospects", *Integrated Manufacturing Systems*, Vol. 11 No. 4, pp. 218–238.

Milgrom, P.R. and Weber, R.J. (1982), "A Theory of Auctions and Competitive Bidding", *Econometrica*, The Econometric Society, Vol. 50 No. 5, p. 1089.

Miroslav, M., Miloš, M., Velimir, Š., Božo, D. and Dorde, L. (2014), "Semantic technologies on the mission: Preventing corruption in public procurement", *Computers in Industry*, Vol. 65 No. 5, pp. 878–890.

Moreno, J.L. (1947), "Contributions of Sociometry to Research Methodology in Sociology", *American Sociological Review*, American Sociological Association, Vol. 12 No. 3, p. 287.

Mori, J., Kajikawa, Y., Kashima, H. and Sakata, I. (2012), “Machine learning approach for finding business partners and building reciprocal relationships”, *Expert Systems with Applications,* Elsevier Ltd, Vol. 39 No. 12, pp. 10402–10407.

Newman, M.E.J. (2004a), “Analysis of weighted networks”, *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics,* Vol. 70 No. 5, p. 9.

Newman, M.E.J. (2004b), “Fast algorithm for detecting community structure in networks”, *Physical Review E,* Vol. 69 No. 6, p. 5.

Newman, M.E.J. and Girvan, M. (2004), “Finding and evaluating community structure in networks”, *Physical Review E,* available at:https://doi.org/10.1103/PhysRevE.69.026113.

Ng, C.S.-P. (2013), “Intention to purchase on social commerce websites across cultures: A cross-regional study”, *Information & Management,* North-Holland, Vol. 50 No. 8, pp. 609–620.

Ng, K.Y.A. and Gujar, G.C. (2009), “The spatial characteristics of inland transport hubs: evidences from Southern India”, *Journal of Transport Geography,* Vol. 17 No. 5, pp. 346–356.

Nguyen, T., ZHOU, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2018a), “Big data analytics in supply chain management: A state-of-the-art literature review”, *Computers & Operations Research,* Pergamon, Vol. 98, pp. 254–264.

Nguyen, T., ZHOU, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2018b), “Big data analytics in supply chain management: A state-of-the-art literature review”, *Computers & Operations Research,* Pergamon, Vol. 98, pp. 254–264.

Nilsson, R., Peña, J.M., Björkegren, J. and Tegnér, J. (2007), “Consistent Feature Selection for Pattern Recognition in Polynomial Time”, *The Journal of Machine Learning Research,* Vol. 8, pp. 589–612.

O’Donovan, P., Leahy, K., Bruton, K. and O’Sullivan, D.T.J. (2015), “Big data in manufacturing: a systematic mapping study", *Journal of Big Data*, Journal of Big Data, Vol. 2 No. 1, p. 20.

Olson, D.L. (2015), "A Review of Supply Chain Data Mining Publications", *Journal of Supply Chain Management Science*, Vol. xx No. xx, pp. 1–13.

Ong, J.B.S., Wang, Z., Goh, R.S.M., Yin, X.F., Xin, X. and Fu, X. (2015), "Understanding Natural Disasters as Risks in Supply Chain Management through Web Data Analysis", *International Journal of Computer and Communication Engineering*, Vol. 4 No. 2, pp. 126–133.

Opresnik, D. and Taisch, M. (2015), "The value of big data in servitization", *International Journal of Production Economics*, Elsevier, Vol. 165, pp. 174–184.

Oracle. (2012), *Big Data for the Enterprise*.

Ovchinnikov, A. (2011), "Revenue and Cost Management for Remanufactured Products", *Production and Operations Management*, John Wiley & Sons, Ltd (10.1111), Vol. 20 No. 6, pp. 824–840.

Oztekin, A., Kizilaslan, R., Freund, S. and Iseri, A. (2016), "A data analytic approach to forecasting daily stock returns in an emerging market", *European Journal of Operational Research*, North-Holland, Vol. 253 No. 3, pp. 697–710.

Padmanabhan, B. and Tuzhilin, A. (2003), "On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities", *Management Science*, Vol. 49 No. 10, pp. 1327–1343.

Pang, G., Casalin, F., Papagiannidis, S., Muyldermans, L. and Tse, Y.K. (2015), "Price determinants for remanufactured electronic products: a case study on eBay UK", *International Journal of Production Research*, Taylor & Francis, Vol. 53 No. 2, pp. 572–589.

Papadopoulos, T., Gunasekaran, A., Dubey, R., Altay, N., Childe, S.J. and Fosso-Wamba, S. (2015), "The role of Big Data in explaining disaster resilience in supply chains for sustainability”, *Journal of Cleaner Production*, Elsevier Ltd, available at:https://doi.org/10.1016/j.jclepro.2016.03.059.

Park, D.-H. and Lee, J. (2008), “eWOM overload and its effect on consumer behavioral intention depending on consumer involvement”, *Electronic Commerce Research and Applications*, Elsevier, Vol. 7 No. 4, pp. 386–398.

Paterson, A.S., O’Gorman, K.D., Leung, D. and MacIntosh, R. (2016), *Research Methods for Accounting and Finance : A Guide to Writing Your Dissertation*, Goodfellow Publishers Ltd.

Philip Chen, C.L. and Zhang, C.Y. (2014), “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data”, *Information Sciences*, Elsevier Inc., Vol. 275, pp. 314–347.

Post and Parcel. (2016), “DHL supporting Chengdu as part of China’s &quot;Belt and Road&quot; initiative | Post &amp; Parcel”, *Post and Parcel*, available at: https://postandparcel.info/73296/news/dhl-supporting-chengdu-as-part-of-chinas-belt-and-road-initiative/ (accessed 15 September 2019).

Prasad, S., Zakaria, R. and Altay, N. (2016), “Big data in humanitarian supply chain networks: a resource dependence perspective”, *Annals of Operations Research*, Springer US, pp. 1–31.

Quinlan, J.R. (1992), “Learning with continuous classes”, *Machine Learning*, Vol. 92, pp. 343–348.

Railwaypro.com. (2017), “Rail transport is cheaper than air but faster than sea”, *Railwaypro.Com*, available at: https://www.railwaypro.com/wp/rail-transport-cheaper-air-faster-sea/ (accessed 13 September 2019).

Rehman, M.H. ur, Chang, V., Batool, A. and Wah, T.Y. (2016), “Big data reduction framework for value creation in sustainable enterprises”, *International Journal of Information Management*, Elsevier Ltd, Vol. 36 No. 6, pp. 917–928.

Roso, V. and Lumsden, K. (2010), “A review of dry ports”, *Maritime Economics and Logistics,* Vol. 12 No. 2, pp. 196–213.

Roso, V., Woxenius, J. and Lumsden, K. (2009), “The dry port concept: connecting container seaports with the hinterland”, *Journal of Transport Geography Journal,* Vol. 17, pp. 338–345.

Rozados, I.V. and Tjahjono, B. (2014), “Big Data Analytics in Supply Chain Management : Trends and Related Research”, *6th International Conference on Operations and Supply Chain Management, Bali, 2014,* No. October.

Rubinov, M. and Sporns, O. (2010), “Complex network measures of brain connectivity: Uses and interpretations”, *NeuroImage,* Vol. 52 No. 3, pp. 1059–1069.

Rudnicki, W.R., Wrzesień, M. and Paja, W. (2015), “All Relevant Feature Selection Methods and Applications”, *Feature Selection for Data and Pattern Recognition,* Springer, Berlin, pp. 11–28.

Ruiz, R., Asgari, N., Farahani, R.Z., Fallah, S. and Hosseini, S. (2018), “OR models in urban service facility location: A critical review of applications and future developments”, *European Journal of Operational Research,* Elsevier B.V., Vol. 276 No. 1, pp. 1–27.

Russell, S.J. and Norvig, P. (2016), *Artificial Intelligence a Modern Approach,* 3rd ed.

Saberi, M., Mahmassani, H.S., Brockmann, D. and Hosseini, A. (2017), “A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks”, *Transportation,* Vol. 44 No. 6, pp. 1383–1402.

Salehan, M. and Kim, D.J. (2016), “Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics”, *Decision Support Systems,* Elsevier B.V., Vol. 81, pp. 30–40.

Samaa Digital. (2017), “‘One Belt, One Road’ brings new opportunities for former Silk Road city of Jiayuguan - Samaa Digital”, *Samaa Digital,* available at:

https://www.samaa.tv/economy/2017/07/one-belt-one-road-brings-new-opportunities-for-former-silk-road-city-of-jiayuguan/ (accessed 14 September 2019).

Sanders, N.R. (2014), *Big Data Driven Supply Chain Management*, Pearson Education, Inc.

Santiago, R. and Lamb, L.C. (2017), “Efficient modularity density heuristics for large graphs”, *European Journal of Operational Research*, Vol. 258 No. 3, pp. 844–865.

Sarker, M.N.I., Hossin, M.A., Yin, X. and Sarkar, M.K. (2018), “One Belt One Road Initiative of China: Implication for Future of Global Development”, *Modern Economy*, Vol. 09 No. 04, pp. 623–638.

Sarwar, F. (2018), “China’s One Belt and One Road: Implications of ‘New Eurasian Land Bridge’ on Global Power Play in the Region”, *NUST Journal of International Peace & Stability*, Vol. 1 No. 2, pp. 131–144.

Saumyadipta, B.L.S.P., Rao, P. and Rao, S.B. (2016), *Big Data Analytics: Methods and Applications*, Springer.

Saunders, M.N.K., Lewis, P. and Thornhill, A. (2019), *Research Methods for Business Students*, 8th ed., Pearson Education Limited.

Schmidt, B., Flannery, P. and DeSantis, M. (2015), “Real-Time Predictive Analytics, Big Data &amp;amp; Energy Market Efficiency: Key to Efficient Markets and Lower Prices for Consumers”, *Applied Mechanics and Materials*, Vol. 704, pp. 453–458.

Seuring, S. (2013), “A review of modeling approaches for sustainable supply chain management”, *Decision Support Systems*, Elsevier B.V., Vol. 54 No. 4, pp. 1513–1520.

Seuring, S. and Müller, M. (2008), “From a literature review to a conceptual framework for sustainable supply chain management”, *Journal of Cleaner Production*, Vol. 16 No. 15, pp. 1699–1710.

Shah, A.K. and Oppenheimer, D.M. (2008), “Heuristics made easy: An effort-reduction framework.”, *Psychological Bulletin*, Vol. 134 No. 2, pp. 207–222.

Shan, Z. and Zhu, Q. (2015), “Camera location for real-time traffic state estimation in urban road network using big GPS data”, *Neurocomputing,* Elsevier, Vol. 169, pp. 134–143.

Sheffi, Y. (2015), “Preparing for Disruptions Through Early Detection Preparing for Disruptions Through Early Detection”, *MIT Sloan Management Review*, Vol. 57 No. 1, pp. 31–42.

Shi, Q. and Abdel-Aty, M. (2015), “Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways”, *Transportation Research Part C: Emerging Technologies,* Elsevier Ltd, Vol. 58, pp. 380–394.

Shu, Y., Ming, L., Cheng, F., Zhang, Z. and Zhao, J. (2015), “Abnormal situation management: Challenges and opportunities in the big data era”, *Computers and Chemical Engineering*, Elsevier Ltd, Vol. 91, pp. 104–113.

Silin, Y., Kapustina, L., Trevisan, I. and Drevalev, A. (2017), “China’s economic interests in the ‘One Belt, One Road’ initiative”, *SHS Web of Conferences,* Vol. 39, p. 01025.

Singh, A., Kumari, S., Malekpoor, H. and Mishra, N. (2018), “Big data cloud computing framework for low carbon supplier selection in the beef supply chain”, *Journal of Cleaner Production,* Elsevier, Vol. 202, pp. 139–149.

Sivamani, S., Kwak, K. and Cho, Y. (2014), “A study on intelligent user-centric logistics service model using ontology”, *Journal of Applied Mathematics,* Vol. 2014, available at:https://doi.org/10.1155/2014/162838.

Song, M.-L., Fisher, R., Wang, J.-L. and Cui, L.-B. (2016), “Environmental performance evaluation with big data: theories and methods”, *Annals of Operations Research,* Springer US, Vol. March, pp. 1–14.

Srinivas, S. and Ravindran, A.R. (2018), “Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework”, *Expert Systems with Applications,* Elsevier Ltd, Vol. 102, pp. 245–261.

St-Aubin, P., Saunier, N. and Miranda-Moreno, L. (2015), “Large-scale automated proactive road safety analysis using video data”, *Transportation Research Part C: Emerging Technologies,* Elsevier Ltd, Vol. 58, pp. 363–379.

Stefanovic, N. (2015), “Collaborative predictive business intelligence model for spare parts inventory replenishment”, *Computer Science and Information Systems,* Vol. 12 No. 3, pp. 911–930.

Stieglitz, S. and Dang-Xuan, L. (2013), “Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior”, *Journal of Management Information Systems,* Routledge , Vol. 29 No. 4, pp. 217–248.

Su, C.-J. and Chen, Y.-A. (2018), “Risk assessment for global supplier selection using text mining”, *Computers & Electrical Engineering,* Pergamon, Vol. 68, pp. 140–155.

Subramanian, R. and Subramanyam, R. (2012), “Key Factors in the Market for Remanufactured Products”, *Manufacturing & Service Operations Management,* INFORMS , Vol. 14 No. 2, pp. 315–326.

Suddaby, R. (2006), “From the Editors: What Grounded Theory is Not”, *Academy of Management Journal,* Academy of Management Briarcliff Manor, NY 10510, Vol. 49 No. 4, pp. 633–642.

Sun, Z., Zheng, J. and Hu, H. (2012), “Finding community structure in spatial maritime shipping networks”, *International Journal of Modern Physics C,* World Scientific Publishing Company, Vol. 23 No. 06, available at:https://doi.org/10.1142/S0129183112500441.

Swami, S. and Khairnar, P.J. (2006), “Optimal normative policies for marketing of products with limited availability”, *Annals of Operations Research,* Kluwer Academic Publishers, Vol. 143 No. 1, pp. 107–121.

Swiss Re Institute. (2017), *China’s Belt & Road Initiative: The Impact on Commercial Insurance in Participating Regions*.

Tan, K.H., Zhan, Y., Ji, G., Ye, F. and Chang, C. (2015), “Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph", *International Journal of Production Economics*, Elsevier, Vol. 165, pp. 223–233.

Tan, M. and Lee, W. (2015), "Evaluation and Improvement of Procurement Process with Data Analytics", *International Journal of Advanced Computer Science and Applications,* Vol. 6 No. 8, pp. 70–80.

Tereyağoğlu, N. (2016), "Market Behavior Towards Remanufactured Products", *Environmentally Responsible Supply Chains,* Springer, Cham, pp. 19–28.

The Ministry of Transport of the People's Republic of China. (2017), *Development of International Dry Port in China,* available at: https://www.unescap.org/sites/default/files/China_EGM Dry Ports_2017.pdf.

Ting, S.L., Tse, Y.K., Ho, G.T.S., Chung, S.H. and Pang, G. (2014), "Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry", *International Journal of Production Economics,* Vol. 152, pp. 200–209.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A. and González, M.C. (2015), "The path most traveled: Travel demand estimation using big data resources", *Transportation Research Part C: Emerging Technologies*, Elsevier Ltd, Vol. 58, pp. 162–177.

Tranfield, D., Denyer, D. and Smart, P. (2003), "Towards a methodology for developing evidence-informed management knowledge by means of systematic review"", *British Journal of Management*, Vol. 14 No. 3, pp. 207–222.

Tsai, C.-W., Lai, C.-F., Chao, H.-C. and Vasilakos, A. V. (2015), "Big data analytics: a survey", *Journal of Big Data*, Springer International Publishing, Vol. 2 No. 1, p. 21.

Tsai, C.-Y. and Huang, S.-H. (2015), "A data mining approach to optimise shelf space allocation in consideration of customer purchase and moving behaviours", *International Journal of Production Research*, Vol. 53 No. 3, p. 850.

Tseng, M.-L., Wu, K.-J., Lim, M.K. and Wong, W.-P. (2019), "Data-driven sustainable supply chain management performance: A hierarchical structure assessment under uncertainties", *Journal of Cleaner Production*, Elsevier, Vol. 227, pp. 760–771.

Tu, W., Li, Q., Fang, Z., Shaw, S., Zhou, B. and Chang, X. (2016), "Optimizing the locations of electric taxi charging stations: A spatial–temporal demand coverage approach", *Transportation Research Part C: Emerging Technologies*, Elsevier Ltd, Vol. 65 No. 3688, pp. 172–189.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012), "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach", *European Journal of Operational Research*, North-Holland, Vol. 218 No. 1, pp. 211–229.

Verma, P., Nandi, A.K. and Sengupta, S. (2018), "Bribery games on interdependent complex networks", *Journal of Theoretical Biology*, Vol. 450, pp. 43–52.

Vieira, A., Dias, L., Santos, M., Pereira, G. and Oliveira, J. (2019), "Supply Chain Hybrid Simulation: From Big Data to Distributions and Approaches Comparison", *Simulation Modelling Practice and Theory*, Elsevier, p. 101956.

Walker, G. and Strathie, A. (2016), "Big data and ergonomics methods: A new paradigm for tackling strategic transport safety risks", *Applied Ergonomics*, Elsevier Ltd, Vol. 53, pp. 298–311.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study", *International Journal of Production Economics*, Elsevier, Vol. 165 No. July, pp. 234–246.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., Fosso Wamba, S., Akter, S., et al. (2015), "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study", *International Journal of Production Economics*, Elsevier, Vol. 165 No. July, pp. 234–246.

Wang, C., Chen, Q. and Huang, R. (2018), “Locating dry ports on a network: a case study on Tianjin Port”, *Maritime Policy and Management*, Vol. 45 No. 1, pp. 71–88.

Wang, C., Li, X., Zhou, X., Wang, A. and Nedjah, N. (2016), “Soft computing in big data intelligent transportation systems”, *Applied Soft Computing*, Vol. 38, pp. 1099–1108.

Wang, G., Gunasekaran, A., Ngai, E.W.T. and Papadopoulos, T. (2016), “Big data analytics in logistics and supply chain management: Certain investigations for research and applications”, *International Journal of Production Economics*, Elsevier, Vol. 176, pp. 98–110.

Wang, J. and Zhang, J. (2016a), “Big data analytics for forecasting cycle time in semiconductor wafer fabrication system”, *International Journal of Production Research*, Vol. 54 No. 23, pp. 7231–7244.

Wang, J. and Zhang, J. (2016b), “Big data analytics for forecasting cycle time in semiconductor wafer fabrication system”, *International Journal of Production Research*, Vol. 7543 No. April, pp. 1–14.

Wang, J., Zhang, L., Duan, L. and Gao, R.X. (2015), “A new paradigm of cloud-based predictive maintenance for intelligent manufacturing”, *Journal of Intelligent Manufacturing*, available at:https://doi.org/10.1007/s10845-015-1066-0.

Wang, S., Wan, J., Zhang, D., Li, D. and Zhang, C. (2015), “Towards smart factory for Industry 4.0: A self-organized multi-agent system with big data based feedback and coordination”, *Computer Networks*, Vol. 101, pp. 158–168.

Wang, W., Li, Z. and Cheng, X. (2018), “Evolution of the global coal trade network: A complex network analysis”, *Resources Policy*, No. 1, available at:https://doi.org/10.1016/j.resourpol.2018.10.005.

Wang, Y. and Hazen, B.T. (2016), “Consumer product knowledge and intention to purchase remanufactured products”, *International Journal of Production Economics*, Elsevier, Vol. 181, pp. 460–469.

Wang, Y., Wang, P., Wang, X. and Liu, X. (2018), “Position synchronization for track geometry inspection data via big-data fusion and incremental learning”, *Transportation Research Part C: Emerging Technologies*, Pergamon, Vol. 93, pp. 544–565.

Wang, Y., Wiegerinck, V., Krikke, H. and Zhang, H. (2013), “Understanding the purchase intention towards remanufactured product in closed-loop supply chains”, *International Journal of Physical Distribution & Logistics Management*, Emerald Group Publishing Limited , Vol. 43 No. 10, pp. 866–888.

Wang, Z., Tu, L., Guo, Z., Yang, L.T. and Huang, B. (2014), “Analysis of user behaviors by mining large network data sets”, *Future Generation Computer Systems*, Elsevier B.V., Vol. 37, pp. 429–437.

Wegner, M. and Küchelhaus, M. (2013), *Big Data in Logistics, DHL Customer Solutions & Innovation*.

Wei, H. and Sheng, Z. (2017), “Dry ports-seaports sustainable logistics network optimization: Considering the environment constraints and the concession cooperation relationships”, *Polish Maritime Research*, Vol. 24 No. S3, pp. 143–151.

Wei, H., Sheng, Z. and Lee, P.T.W. (2018), “The role of dry port in hub-and-spoke network under Belt and Road Initiative”, *Maritime Policy and Management*, Vol. 45 No. 3, pp. 370–387.

Wells, J.D., Valacich, J.S. and Hess, T.J. (2011), “What Signal Are You Sending? How Website Quality Influences Perceptions of Product Quality and Purchase Intentions”, *MIS Quarterly*, Management Information Systems Research Center, University of Minnesota, Vol. 35 No. 2, p. 373.

White, M. (2012), “Digital workplaces: vision and reality”, *Business Information Review*, Vol. 29 No. 4, pp. 205–214.

Witte, P., Wiegmans, B. and Ng, A.K.Y. (2019), “A critical review on the evolution and development of inland port research”, *Journal of Transport Geography*, Elsevier, Vol.

74, pp. 53–61.

Witten, I.H. and Frank, E. (2011), *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, available at:https://doi.org/0120884070, 9780120884070.

Witten, I.H., Frank, E. and Hall, M.A. (2011), *Data Mining : Practical Machine Learning Tools and Techniques,* 3rd ed., Morgan Kaufmann.

Wu, K., Liao, C., Tseng, M., Lim, M.K., Hu, J. and Tan, K. (2017), "Toward sustainability: using big data to explore the decisive attributes of supply chain risks and uncertainties", *Journal of Cleaner Production,* Elsevier Ltd, Vol. 142, pp. 663–676.

Xia, D., Wang, B., Li, H., Li, Y. and Zhang, Z. (2016), "A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting", *Neurocomputing,* Elsevier, Vol. 179 No. ii, pp. 246–26.

Xiande Zhao, Yeung, K., Huang, Q. and Song, X. (2015), "Improving the predictability of business failure of supply chain finance clients by using external big dataset", *Industrial Management & Data Systems,* Vol. 115, pp. 1683–1703.

Xiang, Z., Schwartz, Z., Gerdes, J.H. and Uysal, M. (2015), "What can big data and text analytics tell us about hotel guest experience and satisfaction?", *International Journal of Hospitality Management,* Pergamon, Vol. 44, pp. 120–130.

Xie, Y., Liang, X., Ma, L. and Yan, H. (2017), "Empty container management and coordination in intermodal transport", *European Journal of Operational Research,* Elsevier B.V., Vol. 257 No. 1, pp. 223–232.

Xu, X., Zeng, S. and He, Y. (2017), "The influence of e-services on customer online purchasing behavior toward remanufactured products", *International Journal of Production Economics,* Elsevier, Vol. 187, pp. 113–125.

Yan-Qiu, L. and Hao, W. (2016), "Optimization for service supply network base on the user's delivery time under the background of big data", *2016 Chinese Control and*

Decision Conference (CCDC), Vol. 13, IEEE, pp. 4564–4569.

Yan, W., Xiong, Y., Xiong, Z. and Guo, N. (2015), “Bricks vs. clicks: Which is better for marketing remanufactured products?”, *European Journal of Operational Research*, North-Holland, Vol. 242 No. 2, pp. 434–444.

Yang, D., Pan, K. and Wang, S. (2017), “On service network improvement for shipping liners shipping lines under the one belt one road initiative of China”, *Transportation Research Part E: Logistics and Transportation Review*, pp. 1–14.

Yang, L., Liu, S., Tsoka, S. and Papageorgiou, L.G. (2017), “A regression tree approach using mathematical programming”, *Expert Systems with Applications*, Pergamon, Vol. 78, pp. 347–357.

Yu, L., Zhao, Y., Tang, L. and Yang, Z. (2019), “Online big data-driven oil consumption forecasting with Google trends”, *International Journal of Forecasting*, Elsevier, Vol. 35 No. 1, pp. 213–223.

Yu, M., Fransoo, J.C. and Lee, C.Y. (2018), “Detention decisions for empty containers in the hinterland transportation system”, *Transportation Research Part B: Methodological*, Vol. 110, pp. 188–208.

Yu, R. and Abdel-Aty, M. (2014), “Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data”, *Safety Science*, Elsevier Ltd, Vol. 63, pp. 50–56.

Zangenehpour, S., Miranda-Moreno, L.F. and Saunier, N. (2015), “Automated classification based on video data at intersections with heavy pedestrian and bicycle traffic: Methodology and application”, *Transportation Research Part C: Emerging Technologies*, Elsevier Ltd, Vol. 56, pp. 161–176.

Zeng, Q., Wang, G.W.Y., Qu, C. and Li, K.X. (2017), “Impact of the Carat Canal on the evolution of hub ports under China’s Belt and Road initiative”, *Transportation Research Part E: Logistics and Transportation Review*, available at:https://doi.org/10.1016/j.tre.2017.05.009.

Zhan, Y. and Tan, K.H. (2018), "An analytic infrastructure for harvesting big data to enhance supply chain performance", *European Journal of Operational Research*, North-Holland, available at:https://doi.org/10.1016/J.EJOR.2018.09.018.

Zhang, C., Yao, X. and Zhang, J. (2015), "Abnormal condition monitoring of Workpieces based on RFID for wisdom manufacturing workshops", *Sensors (Switzerland)*, Vol. 15 No. 12, pp. 30165–30186.

Zhang, H., Zhao, L. and Gupta, S. (2018), "The role of online product recommendations on customer decision making and loyalty in social shopping communities", *International Journal of Information Management*, Pergamon, Vol. 38 No. 1, pp. 150–166.

Zhang, J. and Meng, M. (2019), "Bike allocation strategies in a competitive dockless bike sharing market", *Journal of Cleaner Production*, Elsevier Ltd, Vol. 233, pp. 869–879.

Zhang, J., Meng, W., Liu, Q., Jiang, H., Feng, Y. and Wang, G. (2016), "Efficient vehicles path planning algorithm based on taxi GPS big data", *Optik*, Elsevier GmbH., Vol. 127 No. 5, pp. 2579–2585.

Zhang, Y., Ren, S., Liu, Y. and Si, S. (2015), "A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products", *Journal of Cleaner Production*, Elsevier Ltd, available at:https://doi.org/10.1016/j.jclepro.2016.07.123.

Zhang, Y., Zhang, G., Du, W., Wang, J., Ali, E. and Sun, S. (2015), "An optimization method for shopfloor material handling based on real-time and multi-source manufacturing data", *International Journal of Production Economics*, Elsevier, Vol. 165, pp. 282–292.

Zhao, R., Liu, Y., Zhang, N. and Huang, T. (2017), "An optimization model for green supply chain management by using a big data analytic approach", *Journal of Cleaner Production*, Elsevier Ltd, Vol. 142, pp. 1085–1097.

Zhao, Y., Zhang, H., An, L. and Liu, Q. (2018), "Improving the approaches of traffic demand forecasting in the big data era", *Cities*, Pergamon, Vol. 82, pp. 19–26.

Zheng, J., Qi, J., Sun, Z. and Li, F. (2018), “Community structure based global hub location problem in liner shipping”, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 118, pp. 1–19.

Zhong, R.Y., Huang, G.Q., Lan, S., Dai, Q.Y., Chen, X. and Zhang, T. (2015), “A big data approach for logistics trajectory discovery from RFID-enabled production data”, *International Journal of Production Economics*, Vol. 165, pp. 260–272.

Zhong, R.Y., Huang, G.Q., Lan, S., Dai, Q.Y., Zhang, T. and Xu, C. (2015), “A two-level advanced production planning and scheduling model for RFID-enabled ubiquitous manufacturing”, *Advanced Engineering Informatics*, Elsevier Ltd, Vol. 29 No. 4, pp. 799–812.

Zhong, R.Y., Lan, S., Xu, C., Dai, Q. and Huang, G.Q. (2016), “Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing”, *International Journal of Advanced Manufacturing Technology*, Vol. 84 No. 1–4, pp. 5–16.

Zhong, R.Y., Xu, C., Chen, C. and Huang, G.Q. (2015), “Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors”, *International Journal of Production Research*, Vol. 7543 No. September, pp. 1–12.

Zhou, H., Zhang, Y. and Li, J. (2018), “An overlapping community detection algorithm in complex networks based on information theory”, *Data & Knowledge Engineering*, available at:https://doi.org/10.1016/j.dyepig.2010.03.014.

Zhou, L. and Disney, S.M. (2006), “Bullwhip and inventory variance in a closed loop supply chain”, *OR Spectrum*, Springer-Verlag, Vol. 28 No. 1, pp. 127–149.

Zhou, L., Xie, J., Gu, X., Lin, Y., Ieromonachou, P. and Zhang, X. (2016), “Forecasting return of used products for remanufacturing using Graphical Evaluation and Review Technique (GERT)”, *International Journal of Production Economics*, Elsevier, Vol. 181, pp. 315–324.

Zhu, D. (2018), “IOT and big data based cooperative logistical delivery scheduling method and cloud robot system", *Future Generation Computer Systems*, North-Holland, Vol. 86, pp. 709–715.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J. and Nguyen, T. V. (2019), "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach", *International Journal of Production Economics*, Elsevier, Vol. 211, pp. 22–33.

Appendix

Eigenvector centrality scores of 80 Chinese cities derived from 3,110 association rules in Chapte 6.

Chinese supplying cities (rule consequence)	Eigenvector centrality
Anqing	0.002864
Anshan	0.007855
Anyang	0.015709
Baoding	0.030029
Beihai	0.008592
Beijing	0.027165
Cangzhou	0.038252
Changsha	0.009613
Changzhi	0.002495
Changzhou	0.067259
Chaozhou	0.013951
Chengdu	0.05314
Chongqing	0.002495
Dalian	0.173423
Dongguan	0.35768
Foshan	0.685699
Fuzhou	0.272416
Ganzhou	0.002864
Guangzhou	0.728089
Hangzhou	0.276754
Heihe	0.002864
Hengyang	0.027902
Heyuan	0.016815
Honghe	0.011087
Huizhou	0.091107
Huludao	0.002864
Huzhou	0.038252
Jiangmen	0.093603
Jiaozuo	0.002495
Jiaxing	0.178045
Jieyang	0.022827
Jinan	0.094908
Jingdezhen	0.002864

Jinhua	0.26592
Jining	0.040747
Jinzhou	0.066806
Jiujiang	0.007855
Langfang	0.002864
Liaocheng	0.002864
Longnan	0.002495
Nanchang	0.051834
Nanning	0.002864
Nantong	0.06329
Ningbo	0.983638
Puyang	0.013214
Qingdao	0.075398
Qingyang	0.011087
Quanzhou	0.090002
Sanming	0.002864
Shanghai	0.354447
Shangrao	0.002495
Shantou	0.151164
Shaoxing	0.096751
Shenzhen	1
Shijiazhuang	0.019679
Suihua	0.082147
Suzhou	0.130549
Tai'an	0.015709
Taiyuan	0.007855
Taizhou	0.093234
Tangshan	0.014972
Tianjin	0.039642
Weinan	0.021806
Wenzhou	0.350678
Wuhan	0.002864
Wuxi	0.085095
Xiamen	0.098509
Xiangyang	0.002495
Xiaogan	0.010718
Xingtai	0.005359
Xuzhou	0.013582
Yancheng	0.047865
Yangjiang	0.008592
Yantai	0.005728
Yuncheng	0.01931
Zhangzhou	0.098509

Zhanjiang	0.002127
Zhengzhou	0.030766
Zhongshan	0.449608
Zhuzhou	0.016078