4th Student Conference on Operational Research

SCOR 2014, May 2–4, 2014, Nottingham, UK

Edited by
Pedro Crespo Del Granado
Martim Joyce-Moniz
Stefan Ravizza

Editors

Pedro Crespo Del Granado
Lancaster University
p.crespodelgranado@lancaster.ac.uk

Martim Joyce-Moniz
Université Libre de Bruxelles
martim.moniz@ulb.ac.be

Stefan Ravizza
IBM Global Business Services
stefan.ravizza@ch.ibm.com

ACM Classification 1998
G.1.6 Optimization, G.1.10 Applications, G.2.2 Graph Theory, G.2.3 Applications, G.3 Probability and Statistics, H.3.5 Web-based services, I.2.3 Deduction and Theorem Proving

ISBN 978-3-939897-67-5

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at http://www.dagstuhl.de/dagpub/978-3-939897-67-5.

Publication date
August, 2014

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Digital Object Identifier: 10.4230/OASIcs.SCOR.2014.i

OASIcs – OpenAccess Series in Informatics

OASIcs aims at a suitable publication venue to publish peer-reviewed collections of papers emerging from a scientific event. OASIcs volumes are published according to the principle of Open Access, i. e., they are available online and free of charge.

ISSN 2190-6807

www.dagstuhl.de/oasics

Contents

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Preface

We are delighted to present the proceedings of the 4th Student Conference on Operational Research (SCOR 2014). The conference took place from the 2–4 May 2014 at the University of Nottingham (UK) and welcomed PhD students, mainly from European Universities, studying Operational Research, Management Science or a related field. SCOR is a student conference organized by PhD students to showcase their work in a friendly environment while at the same time enabling networking with other researchers in the early stages of their OR careers.

Following the success of previous SCOR conferences, 45 participants attended SCOR 2014, with 20 universities across 9 countries being represented. A number of invited speakers gave exciting plenary talks covering a variety of the latest hot topics in OR. Professor Stewart Robinson, President of the OR Society, gave a talk about how simple models can often deliver insight and good results faster than their more complex counterparts. Delegates took away the important message that the inclusion of every single detail in a model can at some point become detrimental to the modelling process, and that sometimes simple is best! In contrast, we then had a talk from Jacqui Taylor, CEO of Flying Binary and the co-chair of the Analytics Network, a data scientist community supported by the OR Society. Jacqui presented some of her work as a data journalist and gave examples on how Big Data, when presented with tools that allow their exploration, can promote reaction and engagement on important issues. For something a little different at the end of the first day, Piero Vitelli, a freelance trainer at Island 41, gave a fantastic talk on presentation skills. As a new addition to a SCOR conference, the idea behind inviting Piero was to expose delegates to techniques to improve the delivery of their presentations and to provide the opportunity to incorporate some of these techniques when presenting their work at the conference. On the second day, Graham Rand of Lancaster University took us on a journey back to the early days of OR and presented the history of OR arising from the conflicts of war to how its popularity spread in more recent times.

SCOR 2014 consisted of 13 research streams and 40 talks; details about the streams and presentations can be found on the website by downloading the conference booklet `http://www.scor2014.com`. Popular streams included Healthcare, Vehicle Routing problems, Analytics, Heuristics, Optimization, OR in energy, Mathematical Programming, Multicriteria Decision Analysis and Forecasting. Another new addition to the conference this year was a question and answer panel session. Members of the panel consisted of two representatives from industry and two from academia, all of whom had completed a PhD in OR. It was interesting to hear the careers paths of the panel members and delegates were keen to ask about their different experiences.

No conference is complete without an exciting array of social activities! These included dinner on Friday at a local restaurant which was followed by a pub quiz organised by the committee. On Saturday, thanks to the great weather, delegates enjoyed a walk around nearby Wollaton Park. The Hall hosts Nottingham Natural History Museum, but is perhaps more popular for being Bruce Wayne's mansion in the recent Batman movies. Delegates enjoyed a three course meal at the conference dinner, the seating plan of which posed an intriguing optimisation problem for the organising committee; each table should have a balance between UK and European university students, no students of the same university should sit next to each other with the aim of promoting networking, and there should be a balance between students who prefer alcoholic and non-alcoholic drinks. The seating plan proved an optimal solution as many new friendships were made!

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

On behalf of the organising committee, we would like to thank our main sponsor, The OR Society, for enabling us to offer affordable fees to participants and finance this conference proceedings, GOAL (Gower Optimal Algorithms Ltd) for sponsoring the best presentation award, and Prospect Recruitment. In addition, we would like to thank the members of the organising committee, without whose extraordinary efforts, this conference could not have happened. Thanks go to our fellow organizing committee members: Gulmira Khussainova , Franklin Djeumou, Elizabeth Rowse, J. Arturo Castillo-Salazar, and Michael Mortenson.

Lastly, we would like to give a special thanks to all authors who submitted a paper for SCOR 2014, the review process was based on the presentation, quality and originality of the research and there were at least two referees assigned to each paper.

Pedro Crespo Del Granado
Stefan Ravizza
Martim Joyce-Moniz

Organisation

Conference Committee

- Pedro Crespo del Granado, Lancaster University, UK (Program Chair)
- Gulmira Khussainova, University of Nottingham , UK (Local Chair)
- Franklin Djeumou Fomeni, Lancaster University , UK
- Martim Joyce-Moniz, Université Libre de Bruxelles, Belgium
- Michael Mortenson, Loughborough University, UK
- Elizabeth Rowse, Cardiff University, UK
- Arturo Castillo Salazar, University of Nottingham , UK
- Stefan Ravizza, IBM Global Business Services, Switzerland (Chair of the Steering Committee)

Sponsors and Partners

- The OR Society
- A National Taught Course Center in Operational Research (NATCOR)
- Prospect Recruitment
- Gower Optimal Algorithms Ltd. (GOAL)
- Divide by Two

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

List of Authors

Akbari, Vahid
College of Engineering, Koc University
Istanbul, Turkey
vakbarighadkolaei@ku.edu.tr

Bendík, Ján
Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovak Republic
Jan.Bendik@fri.uniza.sk

Ceselli, Alberto
Dipartimento di Informatica, Università degli Studi di Milano
Via Bramante 65 – 26013 Crema (CR), Italy
alberto.ceselli@unimi.it

Chahed, Salma
HSCMG, Faculty of Science and Technology, University of Westminster
115 New Cavendish Street, London, W1W 6UW, UK
mohsen.mesgarpour@my.westminster.ac.uk

Chaussalet, Thierry
HSCMG, Faculty of Science and Technology, University of Westminster
115 New Cavendish Street, London, W1W 6UW, UK
mohsen.mesgarpour@my.westminster.ac.uk

Crone, Sven F.
Department of Management Science, Lancaster University
Lancaster LA1 4YX, United Kingdom
s.crone@lancaster.ac.uk

Dickson, Alex
Department of Economics, University of Strathclyde
130 Rottenrow, Glasgow, G4 0GE, UK
alex.dickson@strath.ac.uk

Gelmini, Alberto
Ricerca sul Sistema Energetico
RSE S.p.A., Via Rubattino 54 – 20134 Milano, Italy
alberto.gelmini@rse-web.it

Katsiampa, Paraskevi
Loughborough University
Loughborough, England
P.Katsiampa@lboro.ac.uk

Kunz, Timo P.
Department of Management Science, Lancaster University
Lancaster LA1 4YX, United Kingdom
t.p.kunz@lancaster.ac.uk

Mesgarpour, Mohsen
HSCMG, Faculty of Science and Technology, University of Westminster
115 New Cavendish Street, London, W1W 6UW, UK
mohsen.mesgarpour@my.westminster.ac.uk

Qazi, Abroon
Department of Management Science, University of Strathclyde
40 George Street, Glasgow, G1 1QE, UK
abroon.qazi@strath.ac.uk

Quigley, John
Department of Management Science, University of Strathclyde
40 George Street, Glasgow, G1 1QE, UK
j.quigley@strath.ac.uk

Righini, Giovanni
Dipartimento di Informatica, Università degli Studi di Milano
Via Bramante 65 – 26013 Crema (CR), Italy
giovanni.righini@unimi.it

Salman, Sibel
College of Engineering, Koc University
Istanbul, Turkey
ssalman@ku.edu.tr

Stutzki, Jan
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, Germany
jan.stutzki@unibw.de

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Taverna, Andrea
Dipartimento di Matematica, Università
degli Studi di Milano
Via Saldini 50 – 20133 Milano, Italy
andrea.taverna@unimi.it

Varlas, Georgios
Department of Business Administration,
University of the Aegean
Michalon 8, Chios 82100, Greece
g.varlas@aegean.gr

Vidalis, Michael
Department of Business Administration,
University of the Aegean
Michalon 8, Chios 82100, Greece
m.vidalis@aegean.gr

A Novel Framework for Quantification of Supply Chain Risks

Abroon Qazi[1], John Quigley[1], and Alex Dickson[2]

1 **Department of Management Science, University of Strathclyde**
40 George Street, Glasgow, G1 1QE, UK
2 **Department of Economics, University of Strathclyde**
130 Rottenrow, Glasgow, G4 0GE, UK
{abroon.qazi,j.quigley,alex.dickson}@strath.ac.uk

Abstract

Supply chain risk management is an active area of research and there is a research gap of exploring established risk quantification techniques in other fields for application in the context of supply chain management. We have developed a novel framework for quantification of supply chain risks that integrates two techniques of Bayesian belief network and Game theory. Bayesian belief network can capture interdependency between risk factors and Game theory can assess risks associated with conflicting incentives of stakeholders within a supply network. We introduce a new node termed 'Game theoretic risks' in Bayesian network that gets its qualitative and quantitative structure from the Game theory based analysis of the existing policies and partnerships within a supply network. We have applied our proposed risk modeling framework on the development project of Boeing 787 aircraft. Two different Bayesian networks have been modeled; one representing the Boeing's perceived supply chain risks and the other depicting real time supply chain risks faced by the company. The qualitative structures of both the models were developed through cognitive maps that were constructed from the facts outlined in a case study. The quantitative parts were populated based on intuition and subsequently updated with the facts. The Bayesian network model incorporating quantification of game theoretic risks provides all the reasons for the delays and financial loss of the project. Furthermore, the proactive strategies identified in various case studies were verified through our model. Such an integrated application of two different quantification techniques in the realm of supply chain risk management bridges the mentioned research gap. Successful application of the framework justifies its potential for further testing in other supply chain risk quantification scenarios.

1998 ACM Subject Classification I.2.3 Deduction and Theorem Proving

Keywords and phrases bayesian belief network, cognitive maps, conflicting incentives, game theory, supply chain risk management

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.1

1 Introduction

There are a number of key debates in the literature of risk focusing on the qualitative and quantitative aspects of risk assessment and therefore, choice of methodology must be given due consideration before its application in the field of supply chain risk management [1]. The application of risk theory to supply chain management is still in its early stages of research and there is requirement of conducting empirical studies of already established models. There is a major research gap of exploring established risk practices in other fields for application in the domain of supply chain risk management [1].

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 1–15

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This research study attempts to bridge the mentioned research gap by introducing a novel approach of combining Game theory and Bayesian belief network techniques. Game theory is used to model situations in which supply chain stakeholders have conflicting incentives while Bayesian belief network is a powerful technique to model the causal interdependency between various risk factors. Such a hybrid risk quantification framework has got its unique benefits as the two approaches complement each other and ignoring risks associated with conflicting incentives results in incorrect modeling of the real time situation.

The framework has been validated against an existing case study [2]. The case study was used to construct cognitive maps followed by modeling of the Bayesian networks. Working papers [3, 4] were consulted for establishing game theoretic modeling. This paper adapts the existing game models to incorporate features of continuous time domain and present value of money. Successful implementation of the framework on a case study advocates its potential for application in other supply chain risk management scenarios.

The concept of supply chain risk management is presented in Section 2. Basics of Bayesian network and Game theory are explained in Sections 3 and 4 respectively. Section 5 delineates the design of a novel framework that captures the dynamics of interacting risk factors. The details of the software are described in Section 6. Section 7 presents results and analysis while the conclusion is drawn in Section 8.

2 Supply Chain Risk Management

There are different perceptions of risk in the context of supply chain risk management. There is no clear distinction between risk and uncertainty in supply chain operations [5]. Risk is attributed to uncertain or unreliable sources that finally contribute to the supply chain disruptions while the uncertainty relates to the matching of fluctuating supply and demand in the processes. There are two important aspects of risk; the probability of risk event and its final impact. Most of the researchers have focused on the negative consequences of impact [6, 7, 8] but there is ambiguity regarding the choice of risk event itself. Similarly, there is no consensus regarding the selection of expected (supplier quality problems) or unexpected (wars, strikes, terrorist attacks) features of risk events. Risk in supply chain management relates to an event with small probability occurring abruptly that incurs major loss to the system. Supply chain risk management is defined as "the management of supply chain risk through coordination or collaboration among the supply chain partners so as to ensure profitability and continuity" [6].

Based on a thorough study by carrying out direct observations of the researchers' output and gathering evidence through surveys of focus groups of researchers, following are the major research gaps in the field of supply chain risk management [9]:

- No clear consensus on the definition of supply chain risk management
- Lack of research on the reactive strategies once the risk event has occurred
- Shortage of empirical research in the field

Based on a thorough review of literature in the fields of risk and supply chain risk management, researchers have recommended following future research directions [1]:

- Lack of understanding of risk in supply chain risk management researchers
- Need for exploring already established risk practices in other fields for application in supply chain risk management
- Requirement of conducting case study based empirical studies in order to determine the current risk management methods used by various supply chains

- Need for developing robust and well-grounded supply chain risk management models that can only be materialized through clear understanding of risk and conducting sufficient number of empirical case studies

The risks can be viewed with respect to three broad perspectives [9]:

- A 'butterfly' depiction of risk that separates underlying causes, actual events and ultimate consequences
- Impact based perception in terms of disruptions and delays
- Network perspective in terms of local-and-global causes and local-and-global effects

3 Bayesian Belief Network

Bayesian belief network (BBN) is a graphical representation of causal relationships between variables and associated uncertainty in the dependency in terms of conditional probabilities [10]. The variables are represented by nodes while an arc (directed between two nodes) represents direct causal relationship. The network must be an acyclic directed graph which means that none of the nodes can be traced back while following the direction of arcs. Each node is provided with a set of conditional probabilities except the root nodes, in order to indicate the influence of parent nodes on the child node.

3.1 Application of BBN in Supply Chain Risk Management

Bayesian belief networks are helpful in benchmarking supplier risk profiles that can be used for the determination of key suppliers having major potential impact on revenues of an organization [11]. The model is designed for determining the supplier's external, operational and network risks. The results of the study can help managers focus on key suppliers based on the maximum Value at Risk (VAR) posed to the company. However, the proposed model seems to be industry specific and therefore, the generality of the Bayesian network is limited in scope. In another similar study, suppliers are benchmarked against their risks based on the sensitivity analysis [12]. However, a major limitation of the study is difficulty to get data from current and potential suppliers in populating the Bayesian network. Bayesian network has also been applied in managing risks associated with large engineering project [13]. A combination of Bayesian network and Total Cost of Ownership methods has also been used for selection of potential suppliers [14]. Bayesian network is suitable for modeling risks in case of buyer's incomplete and uncertain information about the suppliers.

4 Game Theory

Game theory was developed to explain the rationale for taking economic decisions that would not have occurred on the basis of simple cost-benefit analysis. Game theory can help the operations managers take appropriate decisions within a supply chain context [15]. A game in a business setting has following four basic elements [16]:

- The players (supply chain stakeholders)
- The rules of the game (policies, constraints)
- The complete set of actions or decisions for each player
- The outcomes or pay-offs resulting from each set of decisions

		Prisoner 2	
		Confess	Deny
Prisoner 1	Confess	-2, -2	0, -3
	Deny	-3, 0	-1, -1

Figure 1 Prisoner's dilemma.

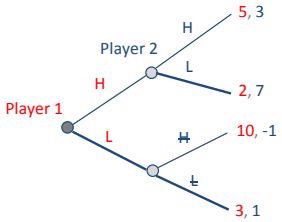


Figure 2 A sequential-move game.

4.1 Simultaneous-Move Games

In simultaneous-move games, each player must take action at the same time or without knowing the moves of other players. The players may have complete information of the pay-offs for each set of decisions. The most popular simultaneous-move game is the prisoner's dilemma where two criminals are apprehended by the Police and asked separately to testify against each other [17]. The game is modeled in Figure 1. Each player can either confess or deny. The first pay-off relates to the row player (Prisoner 1) while the second pay-off corresponds to the column player (Prisoner 2). If both players confess, each gets 2 years of imprisonment. If both deny, each gets 1 year of imprisonment. However, if one confesses and the other denies, the one confessing goes free while the other denying is awarded 3 years of imprisonment. Both the prisoners can be in a better situation by denying but the solution (Nash equilibrium) for this game is both prisoners confessing.

▶ **Definition 1.** *A Nash equilibrium is an action profile a^* with the property that no player i can do better by choosing an action different from $a_i{}^*$, given that every other player j adheres to $a_j{}^*$.* [18]

4.2 Sequential-Move Games

In sequential-move games, players take decisions in sequence. An example of such a game is provided in Figure 2. Solution of such games can be obtained through backward induction. Both the players represent competing industries and need to decide on the price of a product. The strategy of Player 1 is represented by (H, L) while that of Player 2 is given as (H~~H~~, H~~L~~, L~~H~~, L~~L~~). Player 1 has to take the decision first followed by Player 2. By looking at the terminal nodes, it is clear that L~~L~~ is the best strategy for Player 2 and knowing this, Player 1's best strategy is to choose L. Thus, (L, L~~L~~) is the solution (subgame perfect equilibrium) for this game.

▶ **Definition 2.** *A subgame perfect equilibrium is a strategy profile s^* with the property that in no subgame can any player i do better by choosing a strategy different from $s_i{}^*$, given that every other player j adheres to $s_j{}^*$.* [18]

5 A Novel Framework

The novel framework of combining the two techniques of Bayesian belief network and Game theory is shown in Figure 3. The hybrid framework reveals complementary effect of integrating these two modeling methods. Majority of the supply chain quantitative modeling schemes do not consider the risks of misaligned objectives (conflicting incentives) among supply chain partners. Modeling these risks through the Game theory approach and

subsequent incorporation in the Bayesian belief network provide more realistic approach towards quantifying the supply chain risks. Bayesian networks have the advantage of capturing dynamic nature of the interacting risk factors.

Initially, the key risk factors are identified within the supply chain followed by qualitative modeling of the Bayesian network. Game theoretic modeling of the conflicting incentives is carried out through a detailed analysis of available information in the form of policies and/or partnerships. The players are identified and their strategies are established followed by the determination of their pay-offs. Finally, game theoretic analysis is performed and results are incorporated in the Bayesian network in the form of a small network of 'Game theoretic risks'. The 'Game theoretic risks' node is connected to an appropriate impact node and the conditional probability table of the child node is populated based on the game theoretic modeling results.

The entire Bayesian network is populated with conditional probability tables followed by the initial updating. Sensitivity analysis of the game theoretic risks is performed. In case of sensitivity being high, a fair strategy is devised followed by the game theoretic analysis of misaligned objectives. The loop is repeated until the acceptable sensitivity results are obtained. This process is followed by the sensitivity analysis of rest of the risk factors followed by determination of proactive risk mitigation strategies.

6 Software

The cognitive maps have been constructed in Decision Explorer. Bayesian belief networks can be modeled in a number of software including Hugin, AgenaRisk, Graphical Network Interface (GeNie), etc. We used GeNie for modeling and analyzing the networks. The software has been developed by the Decision Systems Laboratory, University of Pittsburg.

7 Analysis and Results

The developed framework is applied on an existing case study concerning the development project of Boeing 787 aircraft [2]. The analysis and results of Game theory and Bayesian network techniques are presented in following sections.

7.1 Game Theoretic Analysis

The discrete time based Game theoretic analysis concerning the development project of Boeing 787 aircraft revealed that there were conflicting incentives among the strategic partners [3]. We have developed our new game models on the basis of same study incorporating features of present value of money and continuous timeframe. Every project comprises two components of costs; direct and indirect costs. Direct costs relate to each task of the project including costs covering labor, material, shipping, etc. Indirect costs do not relate directly to the tasks but these are linked to the project duration. Overhead, delaying penalty, order cancellations and financial losses are some of the examples of indirect costs. A longer task is considered to lower direct costs while a longer project increases indirect costs [19].

The direct cost of a task reduces with the duration representing a convex function as shown in Figure 4. 'X_i' indicates the scheduled timeframe of the task. If either the Boeing or a Tier-1 supplier delays its task, it gets saving represented by 's_i'. In case of expediting the task, there is an associated cost represented by 'c_i'. The indirect cost of a project increases with the project duration representing a convex function as shown in Figure 4. 'X_s' indicates the scheduled timeframe of the suppliers' tasks while 'X_m' indicates the scheduled timeframe

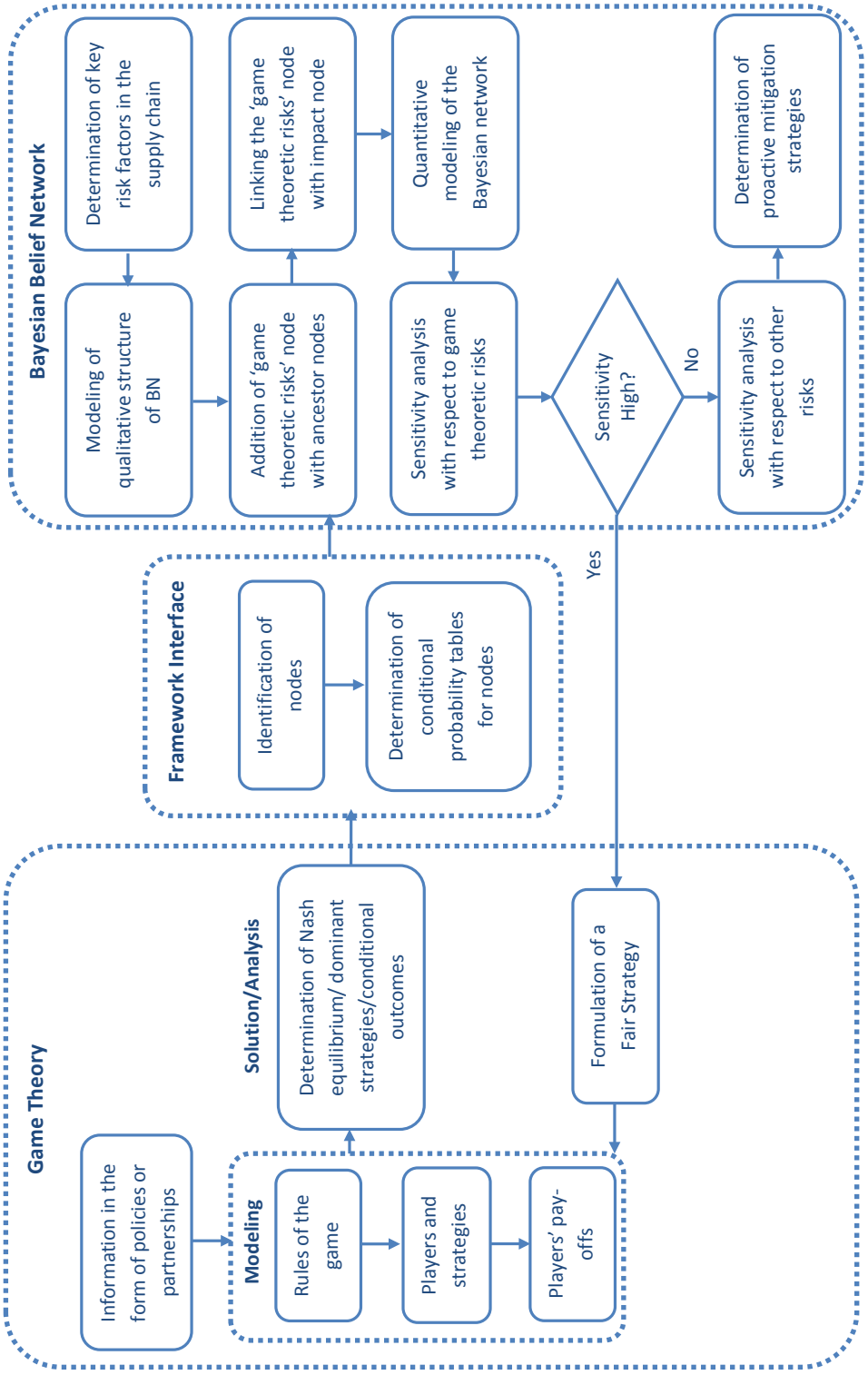


Figure 3 A novel framework for quantification of supply chain risks.

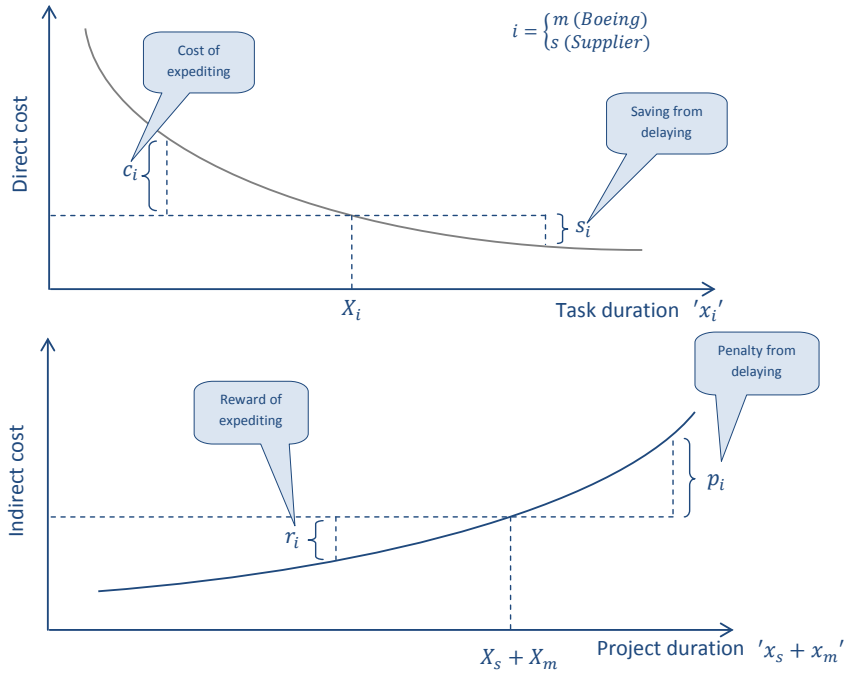


Figure 4 Variation of direct and indirect costs with task and project duration respectively.

of the Boeing's task. If the resultant time of the two tasks exceeds the scheduled time, there is a penalty 'p_i' whereas expediting the project results in the reward 'r_i' for each partner.

It is assumed that all the Tier-1 suppliers perform their tasks in parallel and the overall time taken by the suppliers is determined by the supplier completing its task at the end. After completion of tasks by the suppliers, Boeing performs the final assembling phase.

7.1.1 Strategic Loss Sharing Partnership

The strategic partnership was introduced by Boeing in order to reduce its financial risks. Each firm was supposed to bear the direct and indirect costs whereas final payment was to be made only after the successful culmination of the project. If a firm delays its task and the project gets delayed, all the firms incur additional indirect costs but the delaying firm saves from its direct costs. The firms having completed the tasks in time, are unfairly penalized because of the project delay caused by the delaying firm. As the firms were not made responsible exclusively for their specific actions, this type of partnership resulted into 'moral hazard' [20]. There were misaligned objectives as every supplier would consider the possibility of other partner delaying the respective task and in case of delivering the task in time, the supplier would lose the amount contrary to the delaying suppliers gaining the same. We will present various forms of games in order to analyze the game theoretic perspective of Boeing's partnership.

- In the first form of game, we consider only one Tier-1 supplier and Boeing. Each player can either delay (D) or keep (K) the task schedule. The extensive form of this sequential-move game is presented in Figure 5. Assuming the marginal benefits being lower than the marginal costs for delaying, Boeing's best strategy is to keep the project in time while the supplier also completes the task in time.

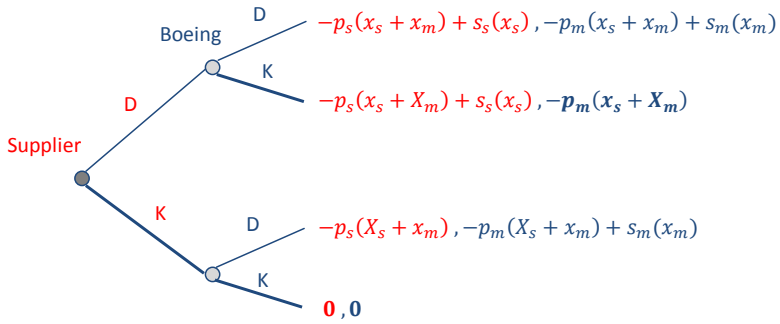
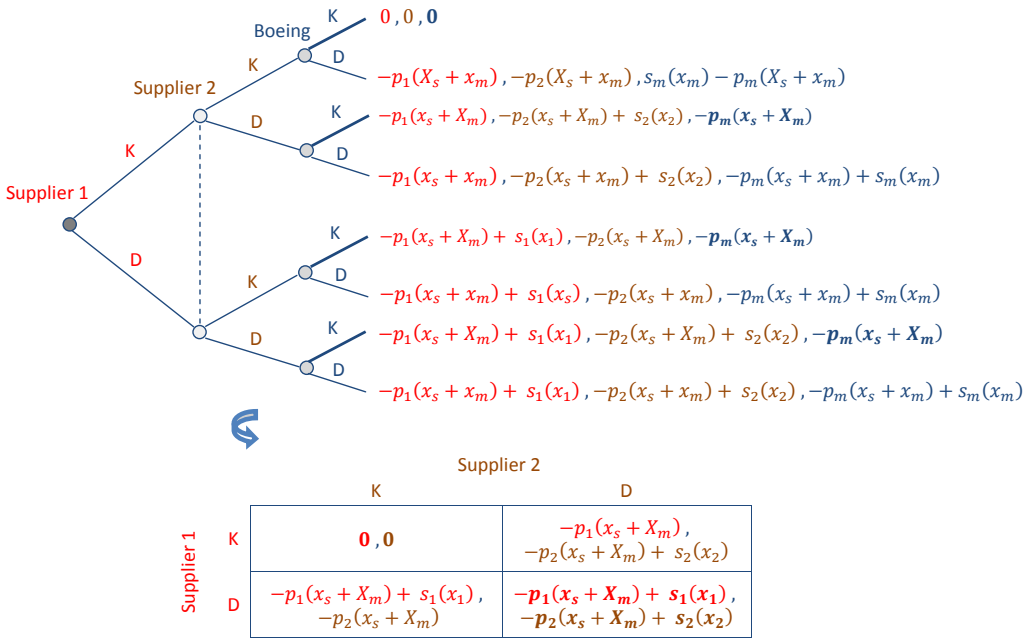


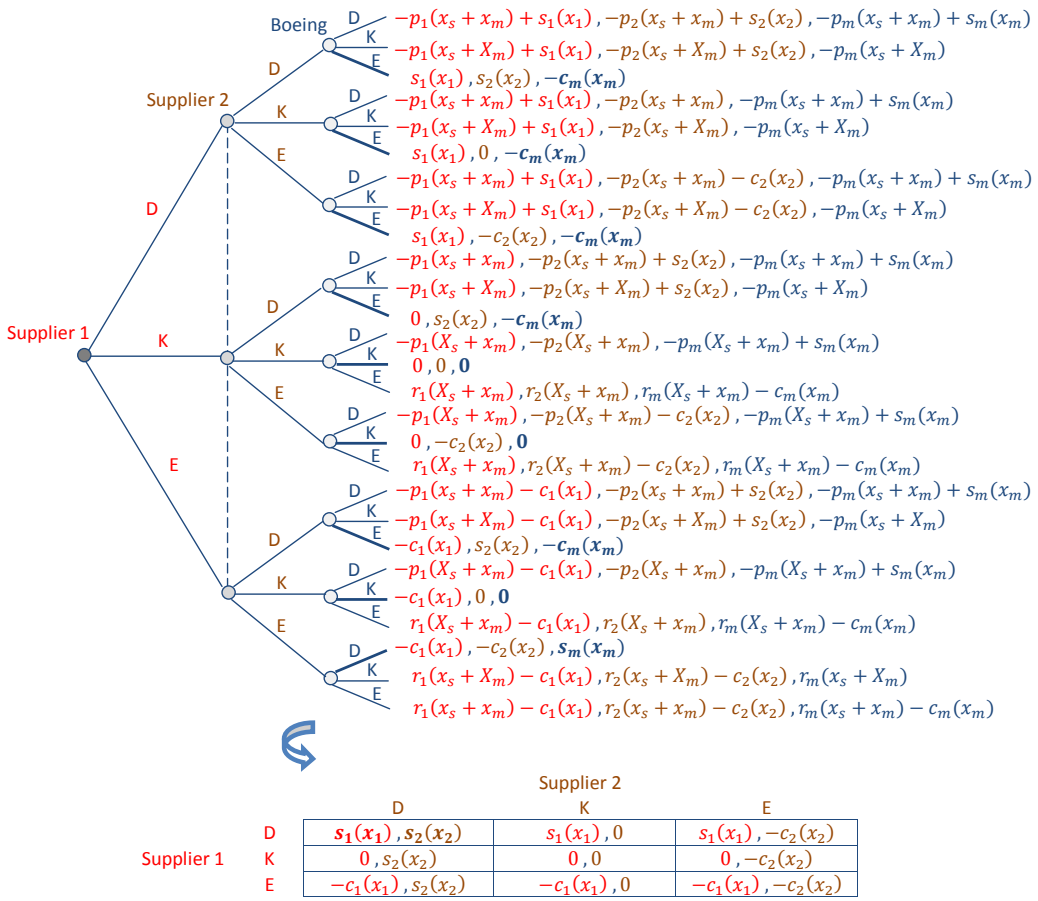
Figure 5 A game between one supplier and Boeing (two actions for each player).



Supplier 1 \ Supplier 2	K	D
K	$\mathbf{0}\,, \mathbf{0}$	$-p_1(x_s + X_m)\,,$ $-p_2(x_s + X_m) + s_2(x_2)$
D	$-p_1(x_s + X_m) + s_1(x_1)\,,$ $-p_2(x_s + X_m)$	$\boldsymbol{-p_1(x_s + X_m) + s_1(x_1)}\,,$ $\boldsymbol{-p_2(x_s + X_m) + s_2(x_2)}$

Figure 6 A game between two suppliers and Boeing (two actions for each player).

- Now, we extend the game to two suppliers and Boeing. The three players' game is presented in Figure 6. The dotted line represents simultaneous game between the two suppliers as they are assumed to perform the tasks in parallel. Assuming the marginal penalty being higher than the marginal saving from delaying, the Boeing's best strategy is to keep the schedule. Having established this fact from the extensive form of the game, the game between suppliers can be modeled in matrix form. From the matrix form, it is clear that the two Nash equilibria are (K,K) and (D,D). Therefore, the strategies of the stakeholders are not aligned and there is a chance that project would be delayed as each supplier considers the possibility of other supplier delaying the task.
- Now, we extend our model to incorporate the option of expediting a task. The extensive form of the game is presented in Figure 7. Based on the technique of backward induction, we can determine the best response of Boeing at each of the terminal nodes. As there is no other sub-game because of the simultaneous game between two suppliers, the solution is determined through matrix form of the game. It is easy to interpret that there is



Supplier 1 \ Supplier 2	D	K	E
D	$s_1(x_1), s_2(x_2)$	$s_1(x_1), 0$	$s_1(x_1), -c_2(x_2)$
K	$0, s_2(x_2)$	$0, 0$	$0, -c_2(x_2)$
E	$-c_1(x_1), s_2(x_2)$	$-c_1(x_1), 0$	$-c_1(x_1), -c_2(x_2)$

Figure 7 A game between two suppliers and Boeing (three actions for each player).

only one subgame perfect equilibrium that is the decision to delay on the part of both the suppliers because Boeing would have to expedite in order to save the heavy penalty resulting from the project delay. The consideration of present value of money does not affect the outcome if it is assumed that Boeing would always prefer completing the task in time even at the cost of expediting.

7.2 Framework Interface

After analyzing the strategic partnership, it is revealed that the partnership engendered misaligned objectives among the stakeholders that finally contributed to the game theoretic risks. As a result of this analysis, three nodes are identified namely 'fair strategy', 'misaligned objectives' and 'game theoretic risks'. The qualitative and quantitative parts of the structure are determined for subsequent incorporation into the Bayesian network.

7.3 Bayesian Network Analysis

The perceived oversimplified cognitive map of the Boeing 787 Project comprised 27 concepts and 38 links. The Bayesian belief network based on the cognitive map is depicted in Figure 8. The model clearly reveals that Boeing was focusing on the opportunities resulting from the

introduction of unproven technology and unconventional supply chain. After the inferencing stage, the probability of development time being high was just 0.09 and that of development cost was 0.22. These favorable results represent the fact that the Boeing management had ignored the interdependency between risk factors and assumed the events of development cost and time being high as unlikely. Contrary to the expectations, the project was delayed by almost 3 years causing major financial penalty to the Boeing.

There were a number of risks associated with the decisions taken by the Boeing management. The cognitive map of the actual supply chain risks comprised 41 concepts and 63 links. A Bayesian network was modeled following the steps outlined in the framework. Three nodes identified earlier as 'fair strategy', 'misaligned objectives' and 'game theoretic risks' were added to the BN. The output of 'game theoretic risks' node was linked to the 'development time' node. The impact of 'game theoretic risks' node on the 'development time' node was quantified based on the game theoretic analysis. The resulting Bayesian network is presented in Figure 9. The unproven technology resulted in major technological risks that further affected the intended outcomes. Outsourcing was considered to be a means of reducing development cost and time; however, it resulted in integration issues as the Tier-1 suppliers were not proficient in selecting their suppliers. Furthermore, the strategic partnership was not a fair strategy as it did not provide due incentives to the stakeholders to keep the schedule in time. This caused increase in the game theoretic risks, being dominant on other factors affecting the development time.

The management involved in the project was lacking expertise in supply chain risk management. The expertise would have provided a guard against all the risks in terms of adopting suitable mitigation strategies. Game theoretic risks are assumed to be independent of the management expertise as the conventional supply chain risk management does not focus on analyzing the risks caused by misaligned objectives of the stakeholders. It also emphasizes the importance of considering unique category of risks within the project risk assessment and the management must possess the ability to apply Game theory to quantify such risks.

The initial updating reveals that the probabilities of development cost and time being high were 0.46 and 0.54 respectively. Different scenarios were generated and the impact of individual risk factor was determined as shown in Table 1. Management expertise was found to be the dominant factor influencing development cost as it lowered its probability being high by 37 % in relation to the case with no management expertise in supply chain risk management.

Game theoretic risks were considered as the dominant factor influencing development time. Introduction of a fair strategy lowered its probability by 26 % in relation to the case with no fair strategy. Once all the facts were entered in the model, the probabilities of development cost and time being high were 0.81 and 0.98 respectively indicating high likelihood of the events. The proactive strategies of ensuring a team with supply chain risk management expertise, devising a fair strategy, negotiating with the labor union and adopting a thorough supplier selection process resulted in the probabilities of development cost and time being high as 0.31 and 0.24 respectively.

7.4 Formulation of a Fair Strategy

The sensitivity analysis of game theoretic risks revealed its major impact on the development time. Therefore, there is requirement of designing a fair strategy in order to reduce the game theoretic risks. The main purpose of a fair strategy is to make each player responsible for one's own deeds [3]. If the suppliers perform their tasks within stipulated time then

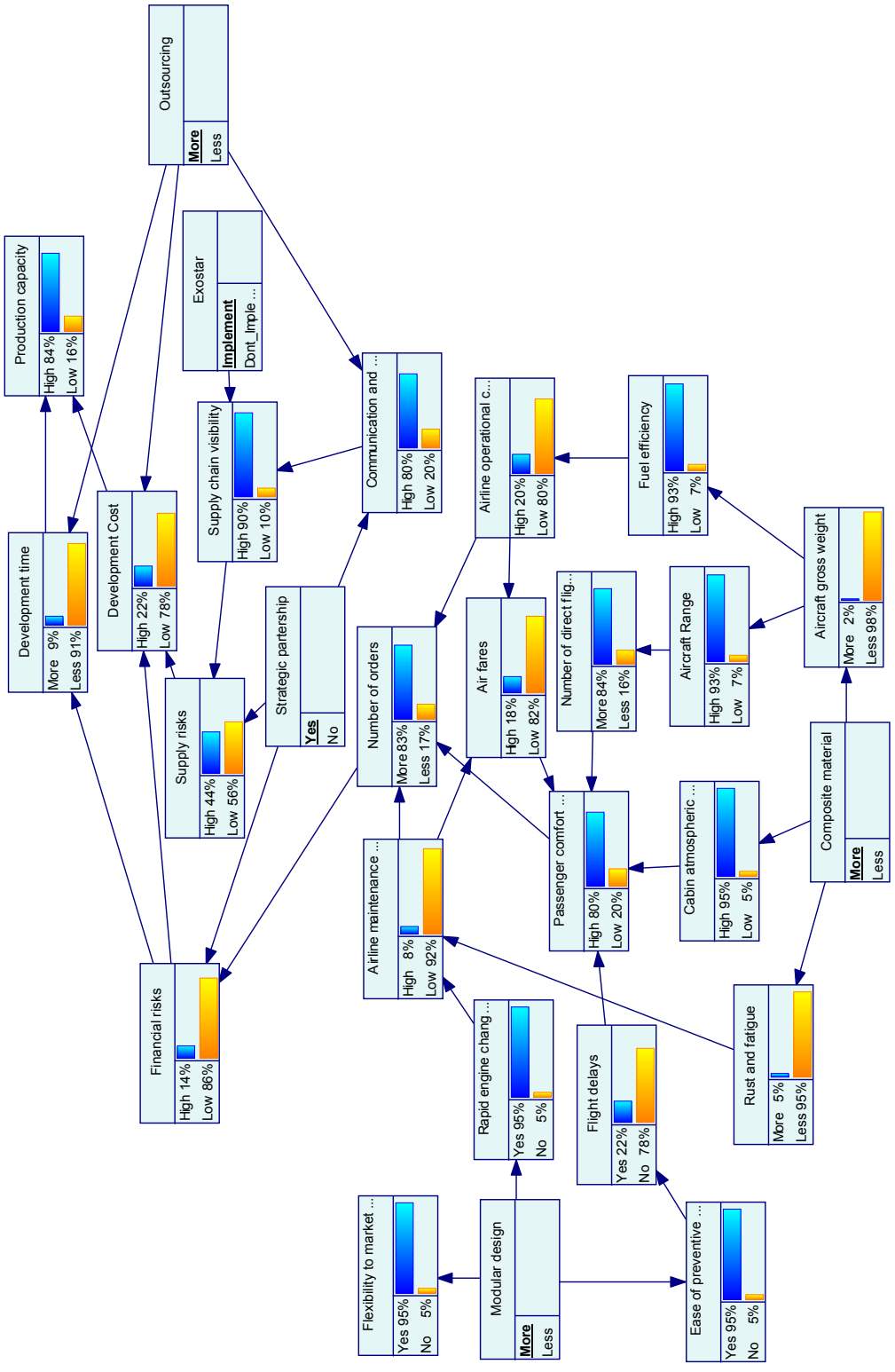


Figure 8 Bayesian belief network based on Boeing's perception.

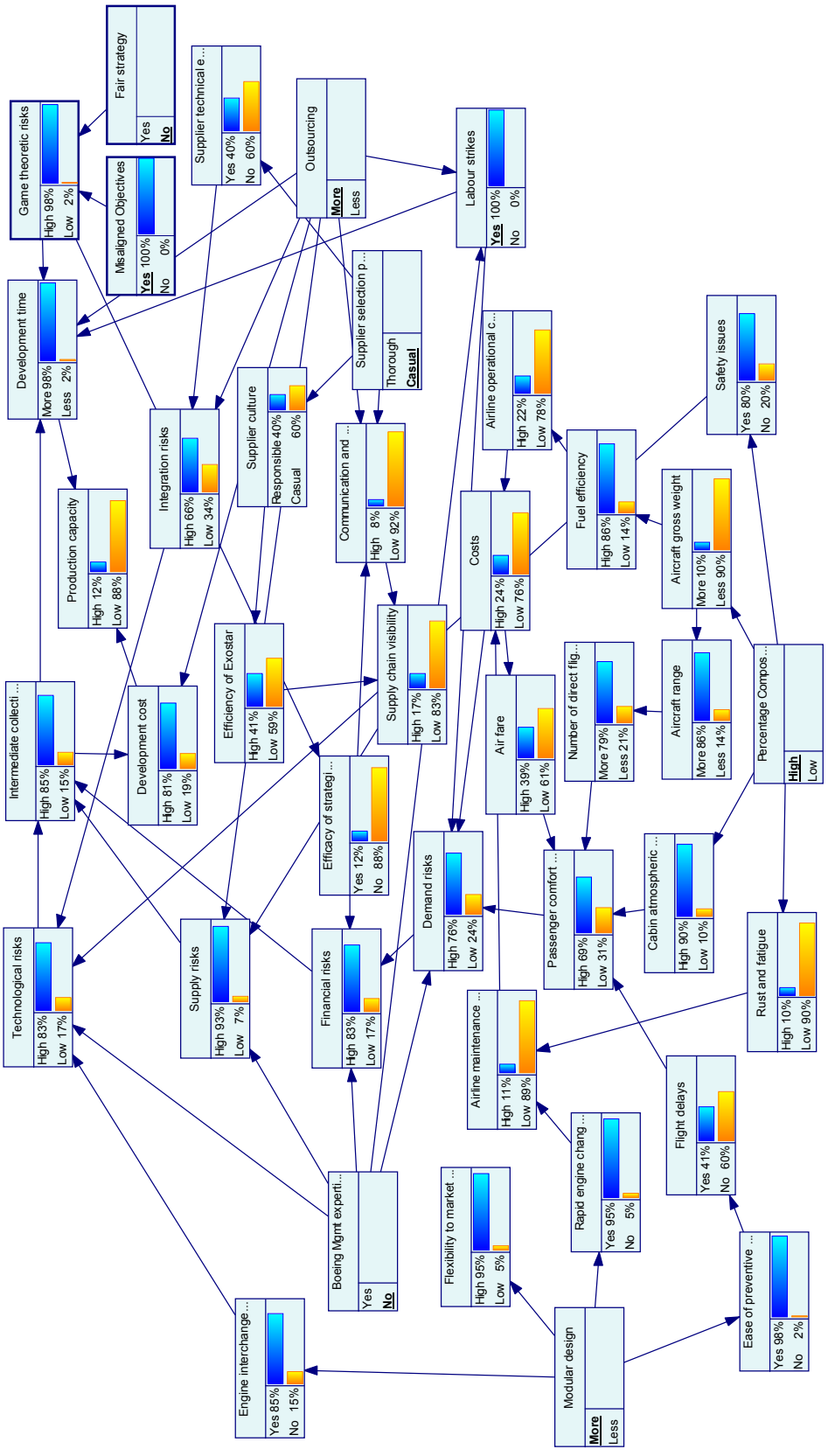


Figure 9 Bayesian belief network based on real time risks.

Table 1 Summary of BBN results.

Scenario	Probability of development time (more)	Probability of development cost (high)
Initial updating	0.54	0.46
Management expertise 'Yes'	0.42	0.28
Management expertise 'No'	0.66	0.65
Fair strategy 'Yes'	0.41	0.44
Fair strategy 'No'	0.67	0.49
Outsourcing 'More'	0.60	0.50
Outsourcing 'Less'	0.48	0.43
Composite material 'More'	0.54	0.48
Composite material 'Less'	0.53	0.45
Modular design 'More'	0.55	0.49
Modular design 'Less'	0.53	0.44
Supplier selection process 'Thorough'	0.53	0.43
Supplier selection process 'Casual'	0.55	0.49
Updating of all facts	0.98	0.81
Fair strategy 'No' and Management expertise 'Yes'	0.93	0.44
Fair strategy 'Yes' and Management expertise 'Yes'	0.33	0.35
Implementation of proactive strategies	0.24	0.31

consequences of any delay on the part of Boeing would be completely compensated by the Boeing and in case of suppliers having expedited their tasks, Boeing would have to pay the reward that did not materialize because of its delay. Similarly, if a supplier is involved in the delay, it will be proportionately penalized for its part of delay. In case of delay on the part of both the suppliers and Boeing, the penalty would be paid fairly.

In the presence of a fair strategy, no partner is incentivized to delay the task; therefore, the project is more likely to be completed in time depending on the other risk factors impacting the delay. After devising the fair strategy, it is revealed that the already existing variables pertaining to game theoretic risks remain same and therefore, there is no requirement of updating the Bayesian network. However, in other situations, the formulation of a policy may necessitate addition of other nodes into the BN requiring some changes as depicted in the framework. After introduction of the fair strategy, the game theoretic risks decrease to the minimum level resulting in lower probability of development time being high.

8 Conclusion

The paper has demonstrated development of a novel framework that combines two complementary techniques of Game theory and Bayesian belief network. The rationale for development of the framework is based on bridging the research gap in the field of supply chain risk management. The developed framework has been successfully applied on the development project of Boeing 787 aircraft. The novel framework captured the dynamics of interacting risk factors. Bayesian belief network is a useful modeling technique for quantification of interdependent risk factors. Game theoretic modeling provides an opportunity to model the risks associated with conflicting incentives among the stakeholders within a

supply network. The Game theoretic results were fed in the Bayesian network as inputs. The results of the study clearly revealed that without mitigating the game theoretic risks, the objective of timely completion of the project was not materialized. Furthermore, lack of management expertise was the major factor contributing to overall costs of the project. The application of this novel risk modeling framework in other supply chain risk projects will help decision makers visualize holistic view of interdependent risk factors and identify key risk factors for establishing proactive risk mitigation strategies.

References

1 Omera Khan and Bernard Burnes. Risk and supply chain management: creating a research agenda. The International Journal of Logistics Management, 18(2):197–216, 2007.

2 Christopher S. Tang, Joshua D. Zimmerman, and James I. Nelson. Managing new product development and supply chain risks: The Boeing 787 case. Supply Chain Forum: an International Journal, 10(2):74–86, 2009.

3 Xu Xin and Yao Zhao. Incentives and coordination in project driven supply chains. 2013. `http://zhao.rutgers.edu/Project-PDSCs.htm` (Accessed 18/02/2014).

4 Yao Zhao. Why 787 slips were inevitable? 2013. `http://zhao.rutgers.edu/Project-PDSCs.htm` (Accessed 18/02/2014).

5 Ou Tang and S. Nurmaya Musa. Identifying risk issues and research advancements in supply chain risk management. International Journal of Production Economics, 133(1):25–34, 2011.

6 Martin Christopher and Hau Lee. Mitigating supply chain risk through improved confidence. International Journal of Physical Distribution and Logistics Management, 34(5):388–396, 2004.

7 Robert E. Spekman and Edward W. Davis. Risky business: expanding the discussion on risk and the extended enterprise. International Journal of Physical Distribution and Logistics Management, 34(5):414–433, 2004.

8 Stephan M. Wagner and Christoph Bode. An empirical examination of supply chain performance along several dimensions of risk. Journal of Business Logistics, 29(1):307–325, 2008.

9 ManMohan S. Sodhi, Byung-Gak Son, and Christopher S. Tang. Researchers' perspectives on supply chain risk management. Production and Operations Management, 21(1):1–13, 2012.

10 J. H. Sigurdsson, L. A. Walls, and J. L. Quigley. Bayesian belief nets for managing expert judgement and modelling reliability. Quality and Reliability Engineering International, 17(3):181–190, 2001.

11 Archie Lockamy III and Kevin McCormack. Modeling supplier risks using bayesian networks. Industrial Management and Data Systems, 112(2):313–333, 2012.

12 Archie Lockamy III. Benchmarking supplier risks using bayesian networks. Benchmarking: An International Journal, 18(3):409–427, 2011.

13 Eunchang Lee, Yongtae Park, and Jong Gye Shin. Large engineering project risk management using a bayesian belief network. Expert Systems with Applications, 36(3, Part 2):5880–5887, 2009.

14 Ibrahim Dogan and Nezir Aydin. Combining bayesian networks and total cost of ownership method for supplier selection analysis. Computers and Industrial Engineering, 61(4):1072–1085, 2011.

15 Heather Lutz, David O. Vang, and William D. Raffield. Using game theory to predict supply chain cooperation. Performance Improvement, 51(3):19–23, 2012.

16 L. Froeb and B. McCann. Managerial Economics: A Problem-Solving Approach. Cengage Learning, 2009.

17 John Nash. Non-cooperative games. The Annals of Mathematics, 54(2):286–295, 1951.

18 J. M. Osborne, 2003. An Introduction to Game Theory. Oxford University Press.

19 Steven Nahmias. 2008. Production and Operations Analysis (McGraw-Hill/Irwin Series Operations and Decision Sciences). McGraw-Hill/Irwin.

20 Bengt Holmström. Moral hazard in teams. Bell J. Econ., 13(2):324–340, 1982.

Multilingual Trend Detection in the Web*

Jan Stutzki

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, Germany
jan.stutzki@unibw.de

Abstract

This paper represents results from our ongoing research project in the foresight area. The goal of the project is to develop web based tools which automatically detect activity and trends regarding given keywords. This knowledge can be used to enable decision makers to react proactively to arising challenges.

As for now we can detect trends worldwide in more than 60 languages and assign these trends accordingly to over 100 national states. To reach this goal we utilize the big search engines as their core competence is to determine the relevance of a document regarding the search query. The search engines allows slicing of the results by language and country.

In the next step we download some of the proposed documents for analysis. Because of the amount of information required we reach the field of Big Data. Therefore an extra effort is made to ensure scalability of the application.

We introduce a new approach to activity and trend detection by combining the data collection and detection methods. To finally detect trends in the gathered data we use data mining methods which allow us to be independent from the language a document is written in. The input of these methods is the text data of the downloaded documents and a specially prepared index structure containing meta data and various other information which accumulate during the collection of the documents.

We show that we can reliably detect trends and activities in highly active topics and discuss future research.

1998 ACM Subject Classification H.3.5 Web-based services

Keywords and phrases Information Retrieval, Web Mining, Trend Detection

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.16

1 Introduction

The Web is the number one information source of our time. Almost all the knowledge of todays world is available online. Current news gets published almost in real time. Developments of events and what people are interested in is mirrored by the content of popular and relevant online documents. For a long time now the Web is in an transition from a tool used by scientists to an everyday commodity used by anyone [1]. While most of the web pages are still written in English the part which is not is growing in absolute and relative size [2].

With the global community expressing its interests and worries online (see Figure 1) it seems appropriate to develop a system which is able to track certain topics and their developments trough published Web documents. It is required to have an reliable detection mechanism for activities and trends available to be able to focus on certain activities and keep track of them over time.

* This work was partially supported by the Planungsamt der Bundeswehr.

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 16–24

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

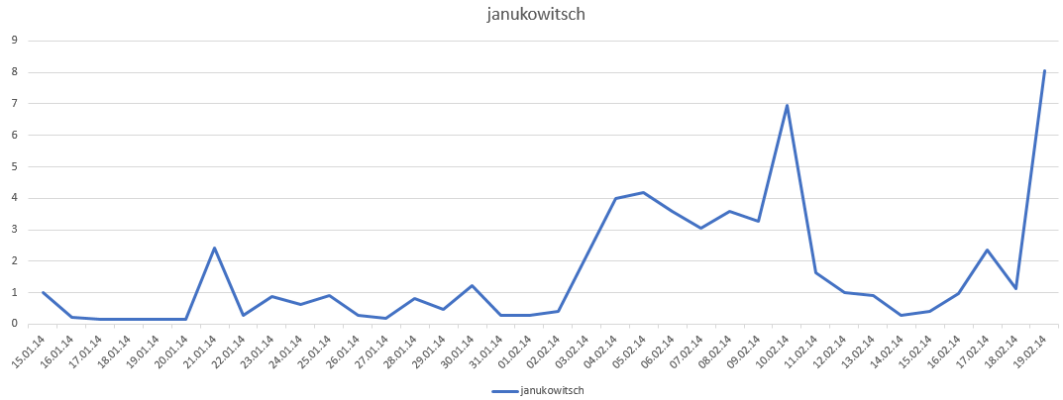


Figure 1 Relative term frequency of "janukowitsch" at the brink of the Ukraine unrest.

Our work focuses on activity and trend detection in the Web. We develop a concept which enables the collection and analysis of potentially relevant data in a scalable and robust way. We propose several methods which are the focus of this paper for automated activity and trend detection in the Web using web services to do the relevance rating for us. We analyze these methods in regard of their ability to detect either and check their performance for different settings. The setting differs in focus of the application and availability of data. To evaluate the methods and as a proof of the concept we developed a prototype (see Figure 2). The web based application is easy to use and provides quick access to the results of most of the proposed methods.

The paper is structured as follows. In Section 2 we explain how data is acquired and what kind of information is collected in what fashion and explain our overall concept for a new approach to activity and trend detection. Section 3 discusses the definitions of trends and activity and introduces the methods which we propose for detection and tracking. The analysis for each method covers its strengths and weaknesses. The applicability of the method for trend detection and activity measurement is also covered. Section 4 contains the conclusion and gives an outlook for further research.

2 Concept

This section focuses on what information is required by the user of the proposed system and which information is gathered automatically. Our concept for data retrieval and analysis consists of six steps:

1. defining search terms
2. translating search terms
3. selection country and language combinations
4. querying search engines for documents
5. downloading the documents
6. analyzing the documents

The task of presenting the results of the documents is regarded as an extra function which is independent of the data collection and analysis process. The idea behind our concept is that experts define the general area in which to look for trends. An expert might be tasked to find new trends and/or centers of activity and developments for a certain topic. Without any further knowledge it may be hard to know where to start.

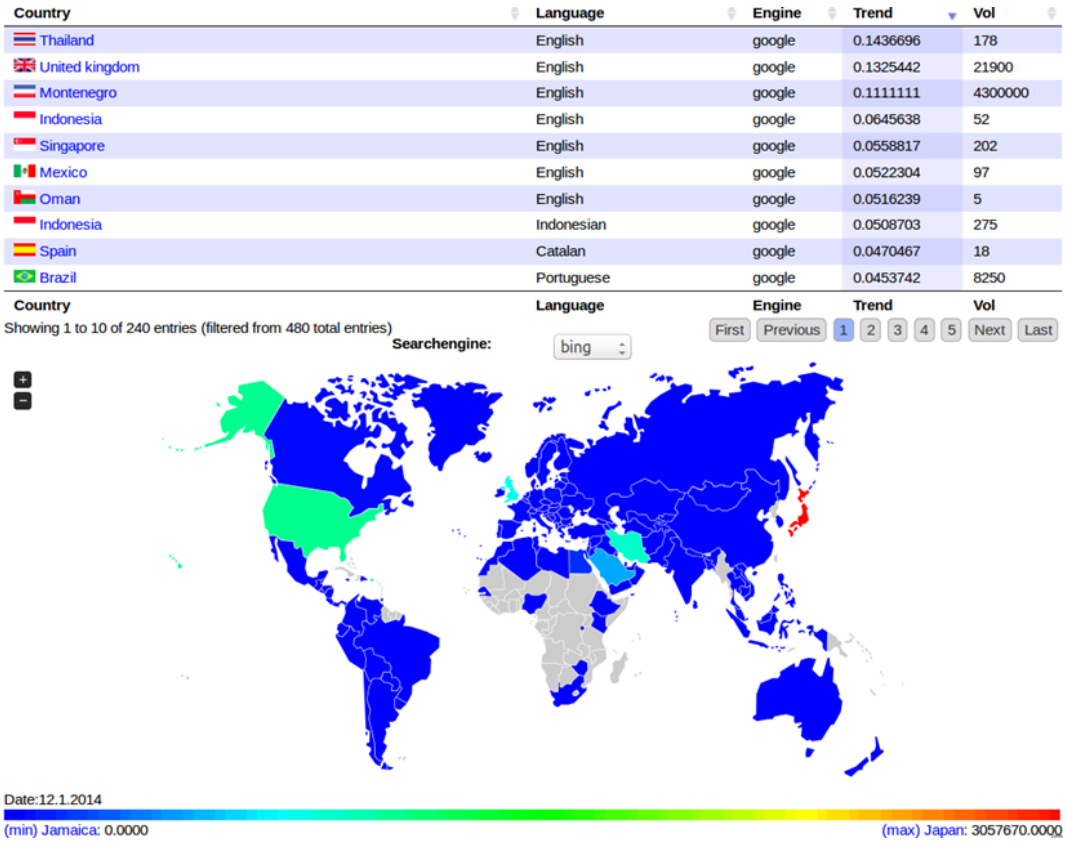


Figure 2 Screen shot of the prototype.

The expert defines a few topic specific search terms in his or her own language and uses a web interface to transfer them to a web based tool. The tool can automatically translate the search terms in the most used languages and allows the user to adapt the automatic translation in case of mistakes [5].

In the next step the user selects which language should be considered for which country. This saves a lot of resources as not all combinations are required. By default for each country the official languages (if available) and English will be preselected as almost 70 % of the Web is in English [3] and English therefore can be assumed as an universally used language independent of the country.

Using the county and language combinations together with the translated search terms the system starts to query one or more search engines for URLs of relevant documents. As Google and Bing do punish scraping [7] we use the official APIs which for Google is Custom Search Engine (CSE) and Bing Azure Websearch. Scraping retrieves the resulting URLs from the official HTML result pages of the search engine by simulating a user and a browser via software. Web search engines disallow the use of scraping because software does not click on advertisements or leaves personal information so there is no source for revenue for the service provider. The punishment for scraping is often done by excluding the client from the service which would be detrimental to our goal of retrieving URLs in fixed intervals. We refer to one run of downloading documents as one crawl and the component doing the downloading the crawler.

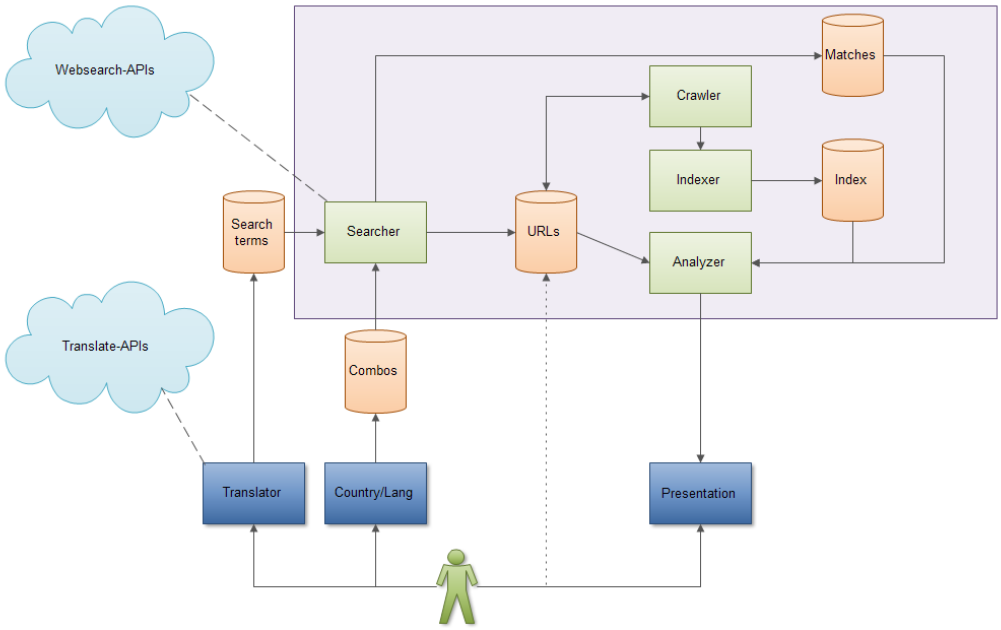


Figure 3 System structure.

The fixed interval between crawls is termed the crawl interval. The shorter the crawl interval is, the less time the data has to change. For very static topics a long search interval might be sufficient while for active topics like news a daily search interval is already at risk of loosing developments. For the analysis of the results of the crawls it is important that the interval between crawls is equidistant.

The crawler will then start to download the documents indicated by the URLs from the search engines. Not all documents are available to any crawler [4] and not every document is actually usable for trend or activity detection. For this reason the crawler has to be exceptionally robust. For analysis we have several information sources. We have the data from the search engines which includes a list of URLs and an estimate of overall matches to the query. We also have the results of multiple crawls and the time when each crawl was done. From the crawls we can deduct several data structures we can use for further analysis. The system is shown in Figure 3. The whole violet box is operating unsupervised and is responsible to perform the crawls in the predefined intervals.

3 Methods

We analyzed several of the following methods and evaluated them in regards to their ability for activity measurement and trend detection:

- Relative Term Frequency
- Estimated Matches
- Page Updates
- New Sources

All methods have in common that results can only be seen relative to previous results. Informally we define trend as a directed activity. An activity is a change between t_0 and t_1 which is detectable within the data we collect. We are particularly interested in activity created by humans. So our methods have to filter the changes in the data for noise and

for changes actually resulting from human behavior. With this trend definition everything done by humans would qualify as a trend. The direction of an activity can be deducted by an increase or decrease of an inspected parameter. While there is no problem with seeing everything as a trend we use "reach" as a property of a trend which allows us to filter for trends which occur on a given percentage of documents. One objective unit for a trend we propose is a term as terms pose a standard used by most documents.
Activity is more generally the deviation between two measurements. This definition allows us to utilize all activity detection methods to some degree for trend detection as a trend is a directed activity. On the other hand activity measurement methods do not necessarily allow us to detect trends as the direction (e. g. rising/falling) component might be missing or carries no value. We use the activity to measure if a topic is still active e. g. there is still research done or has become static.

During our research we also considered synonyms and stemming to be relevant for trend detection as we develop some methods which are text based. We did not regard synonyms as they are heavily context dependent and our index structure currently does not support this. This might change with a positional index but the challenge to extract or collapse the right synonyms for the supported languages remains. Also there are few sources for comprehensive data about multilingual synonyms. After experiments with stemming we decided against it as too much relevant information is lost during stemming (e. g. gender for job descriptions in German) so that a trend might get lost among the other terms which share the same stemmed form while only slightly reducing the dictionary size of the inspected languages. Experiments with a German dictionary and various sources lead to an approximate reduction of the dictionary by only 15 %. With the same reasoning we decided against the usage of character folding. While it is certainly reasonable to use it in a information retrieval context we concluded that its effects are detrimental to trend detection.

For each method we evaluate what data is required and if it is available. Furthermore we consider the dimensions inspected by the methods and a fitting way of visualization and estimate how complex a method is to run on current standard hardware (Intel i5, 8 GB RAM). Because we target documents in several languages and therefore our methods have to work with a multitude of languages we disregard semantic approaches which are usually language specific and focus on statistical analysis of the text corpus at hand.

3.1 Relative Term Frequency

We define the trend of a term t_w as the incline of the trend line of data points of the relative term frequency $F = (f(w)_0, \ldots, f(w)_i)$. The term w has to be part of each data point d_i inspected. The relative frequency $f(w)_i$ is calculated $f(w_i) = |w_i|/|w_0|$ using the frequency of t_0 as reference. The trend line is fitted with least square linear regression to the whole time series inspected. Using relative term frequency allows us to work without stop lists.

This method requires a time frame as an additional input as terms might not be present in the documents crawled at a given time. While the absence of a term could be used as an indicator it saves memory and improves the performance when terms which are not present from t_0 to t_i are ignored.

The relative frequency can be used as an indicator for activity as any change in the relative frequency can be tracked back to change in the underlying text. A high activity is expressed by huge changes in the overall relative term frequency (see Figure 1). To get good results it is necessary to look at several, if possible, all terms.

For trend detection this method is recommendable as each relative term frequency time line can be understood as a trend indicator. Paired with other metrics as for example "reach"

it can indicate which terms are currently rising over proportionately in frequency which we interpret as interest of the humans behind the machines.

In combination with least square linear regression a long term trend can be derived from the data and enable the user to focus on a few exceptional instances of terms for further inspection. Because this method needs to tokenize a text in order to extract terms, problems with languages which do not have an explicit word delimiter arise (e. g. Chinese and Japanese). For other languages this method exceeds expectations and will be part of future research to improve the trend filter and make exceptional trends more obvious. The problem with the tokenizer is an area of active research [6].

We can use this method for the dimensions: language, country, time and text corpus. Depending on the flexibility of the time frame and the inspected dimensions the computational complexity of the method is rather high as the tokenizer needs preprocessed input where the documents are sanitized and cleaned of any markup. While this can be precalculated if the time frame is fixed, non fixed time frames need to be calculated on demand. This can be supported by index structures which are generated after a crawl is finished, but for this to be feasible the dimensions and aggregation level needs to be defined in advance.

Visualization is done on a per term basis. For each term a graph is generated for the inspected time frame. The terms are presented as a sortable list. The list is sortable by term, reach and incline of the trend line.

3.2 Estimated Matches

The number of "Estimated Matches" is acquired during a crawl. It is the number of estimated results as returned by a search engine as part of the reply to a specific query. As all queries are country and language specific we can assign the number of estimated matches to a country and a language.

Depending on the target variable, the results of one crawl may be sufficient. Questions like "where in the world are the most matching documents located" can be answered. Because all languages of a country are treated separately we aggregate the number of estimated documents for each country over all corresponding languages. If a trend is to be detected, at least two crawls are required to retrieve two data points which then can be used to estimate a trend. The estimate number of matches for a given query provided by the search engines varies (see Figure 4). Especially Bing offers unreliable results. In order to give an informed trend many more than two data points are required. Google performs slightly more stable but is still behind expectations. The search engines offer no explanation on why the estimates vary. It can be assumed that different versions of the index are used for the estimate.

For activity detection it is required to have at least two observations. While the number of estimated results from one observation can be used to derive a general feeling for the importance of the matter in question we need two observations to get an idea in which direction possible activity is heading. For large numbers Google and Bing provide very volatile results (50 %–70 % change in a matter of hours) which obviously does not reflect the actual state of the Web. The estimated matches proved to be more stable for a smaller result set. Therefore we suggest this method in regards to activity measurement only for topics which yield a small number (a few hundreds to thousands) of results.

As far as trend detection is concerned we would recommend against using this method as the number of results varies too much and a fluctuation might induce a trend which is not actually there though it seems to stabilize if there are enough measurements done. With this method we consequently regard the dimensions time, country and language. We could also add search engine as another dimension. For visualization a world map and a slider for the

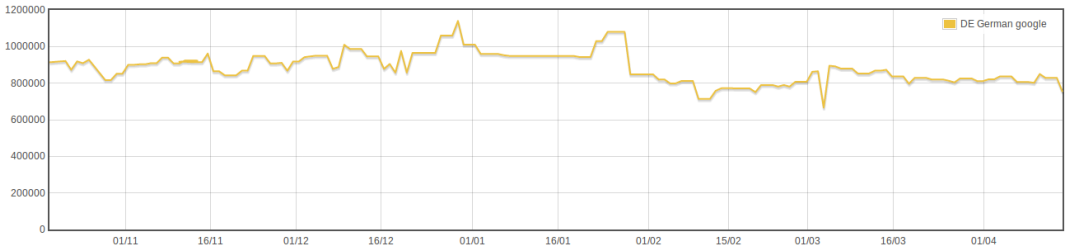


Figure 4 Estimated Matches by Google for "news" in German.

time dimension would be sufficient. The time series can be displayed as graphs. The basis data needed for this method is a byproduct of each query. As almost no processing is taking place this method is easily implemented and has low hardware requirements.

More data is required and further research has to be done to give estimates how much data is necessary for a good trend detection and activity measurement.

3.3 Page Updates

This method requires at least the data of two crawls to work. The method extracts all URLs which are part of the crawls c_{t_0} and c_{t_1}. As the documents were retrieved by the crawler and are stored locally at t_0 and at t_1 we can compare these two versions of the document. If they are dissimilar we assume that an update of either the content or the document structure was made. As the absolute number of updated documents varies as the index of the search engines changes and the ranking of the documents get reevaluated we use the relative update rate of all documents which are part of the intersection of c_{t_0} and c_{t_1}.

Activity is detected easily as an updated document indicates that either by a user, web master or at least by an automated system the effort was undertaken to change content. Counting any change as activity is a generalization we chose on purpose as it drastically improves speed of analysis compared to other approaches and avoids the problem of quantifying the degree "change" of documents. The downside is that things like an automatically generated time stamp on a page would increase the activity rating besides no activity taking place. While this could be fixed with a constant pool of documents where the effects of a rogue document would be canceled out over time we decided to take the risk because we value the relevance rating of the search engines over the occasional misinterpretation of activity.

In regards to trend detection this method is of limited use. A direction can be deducted by the increase or decrease of the activity over time. But currently we can not extract more specific information about the underlying forces of a trend from this method.

First experiments with this method have shown that it provides rather stable results and is opposed to e. g. Estimated Matches not prone to unreliable search engine data. A comparison of two topics showed that "daily news" has a 95 % update rate while the more stable "geoengineering" only has a 25 % update rate aggregated over the whole topic. This method also works independently from the number of documents. The recommended amount of documents for this method is still subject of ongoing research.

The dimensions inspected by this method are time, country, language and search engine. Various aggregations e. g. all languages of a country are possible. Changes of the results by different ways of aggregating are also subject of ongoing research.

Visualization is done by a chart and a world map. The computational complexity depends on the way differences between document versions are detected and treated. A more complex

analysis can be costly. Our approach via a hash allows us to keep the cost in terms of computing time low.

3.4 New Sources

A basic requirement is the results of two crawls c_{t_0} and c_{t_1}. By comparing the URLs provided by the search engines we look for sources which were previously not part of the crawl. The assumption is that if a search engine adds or removes URLs to the result set of a specific query something must have happened to change the relevance rating of this particular or following URLs. Therefore we assume that some kind of activity has happened which results in the changed URLs. We use the search results because big search engines are more capable to analyze a large part of the Web and have a tested infrastructure.

Naturally this method is primarily developed with activity measurement in mind. We can show that a topic like "geoengineering" behaves differently from "news" but due to the way modern search engines handle their index structure this method suffers from the same symptoms as the Estimated Matches method. Search engine provider usually operate with several index structures in parallel. A "current" index is used to answer queries while the next index is built in the background. Adaptations of the current index are done by lists which contain deleted URLs [3]. Currently it is impossible for us to distinguish between actual activity and the switching to a new index at the side of the search engine provider. A high influx of new URLs might suggest a new index structure and in the case of successive large changes activity might be deducted but we can not cancel out the possibility of the reply to the search queries coming from different data centers or index structures. We have to conduct further research to see how this method performs in the long run and how we can detect which index version is delivering the URLs so we can compare changes to the same index.

For many observation this method can be used for trend detection. In the current state the results are too imprecise and fuzzy to detect a direction of the activity.

This method is also working on query basis which allows us to inspect the dimensions time, country, language and search engine. Similar to Page Updates aggregation can be used to get more general information e. g. about the development in a particular country.

For visualization we currently only use a chart as the relationship between available documents and index rebuilding and effect on the result set is not yet clear so a country-by-country comparison does not to seem feasible yet.

4 Conclusions & Outlook

In this paper we presented a new approach to activity monitoring and trend detection in the Web. The new approach consists of a a new concept to acquire data and a number of methods for activity and trend detection. We explained the concept for data acquisition and used an fully implemented prototype to prove its feasibility. Then we continued to show what methods can be applied to the data and which conclusions can be drawn from each method using examples and gathered data.

Activity measurement and trend detection is tightly linked. To develop better methods for trend detection we need the ability to detect activity so that we can research the various causes of the activity. In order to improve our activity measurement methods we need to collect more data and research the relationships of the various dimensions we deal with e. g. search engine.

We plan to use the developed methods to analyze the performance of various search engines in regards to index variance and refresh rate. Also we need to evaluate the stability of the results of the search engine providers so that we can give a qualified suggestion which providers are best fit for our methods.

In regards to trend detection we research the average live span of trends in various topics. Also a field of our research is the clustering of trend terms via correlation analysis and spatial analysis based on the downloaded documents.

Further technologies which we would like to include in future work are positional indexes. Positional indexes enables the detection of terms which consist of more than one token (e. g. New York) by searching for combinations of terms which appear close to each other (n-grams). The proposed text-based methods for activity and trend detection are unaffected by a positional index as an abstraction layer can be built to keep the input format unchanged.

We plan to validate the methods using the Reuters Information Retrieval Text Research Collections though there is no such thing as a gold standard for trend detection on unstructured data.

References

1 Wangberg, Silje C., et al. *Relations between Internet use, socio-economic status (SES), social support and subjective health.* Health promotion international 23.1 (2008): 70–77.

2 Miniwatts Marketing Group. *Internet World Stats.* 31 May 2011, accessed 16 April 2014

3 Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich. *Introduction to information retrieval.* Cambridge university press Cambridge, USA, 2008

4 Sun, Yang and Zhuang, Ziming and Giles, C. Lee. *A large-scale study of robots. txt.* Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

5 Papineni, Kishore, et al. *BLEU: a method for automatic evaluation of machine translation.* Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

6 Fung, Pascale. Extracting key terms from Chinese and Japanese texts. Computer Processing of Oriental Languages 12.1 (1998): 99–121.

7 McCown, Frank and Nelson, Michael L. *Search engines and their public interfaces: which apis are the most synchronized?* Proceedings of the 16th international conference on World Wide Web. ACM, 2007

Solving the p-median location problem with the Erlenkotter approach in public service system design

Ján Bendík

Faculty of Management Science and Informatics, University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovak Republic
Jan.Bendik@fri.uniza.sk

Abstract

This work deals with the problem of designing an optimal structure of a public service system. The problem can be often formulated as a weighted p-median problem. Real instances of the problem are characterized by big numbers of possible service center locations, which can take the value of several hundreds or thousands. The optimal solution can be obtained by the universal IP solvers only for smaller instances of the problem. The universal IP solvers are very time-consuming and often fail when solving a large instance. Our approach to the problem is based on the Erlenkotter procedure for solving of the uncapacitated facility location problem and on the Lagrangean relaxation of the constraint which limits number of the located center. The suggested approach finds the optimal solution in most of the studied instances. The quality and the feasibility of the resulting solutions of the suggested approach depends on the setting of the Lagrangean multiplier. A suitable value of the multiplier can be obtained by a bisection algorithm. The resulting multiplier cannot guarantee an optimal solution, but provides a near-to-optimal solution and a lower bound. If our approach does not obtain the optimal solution, then a heuristic improves the near-to-optimal solution. The resulting solution of our approach and the optimal solution obtained by the universal IP solver XPRESS-IVE are compared in the computational time and the quality of solutions.

1998 ACM Subject Classification G.1.6 Optimization

Keywords and phrases p-median location problem, Erlenkotter's approach, Lagrangean relaxation

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.25

1 Introduction

The p-median location problem has became one of the most well-known and studied problems in the field of facility location. Solving the p-median location problem in the public service system design is the NP-hard problem [6, 7]. The public service system structure is formed by deployment of limited number of the service centers and the associated objective is to minimize costs. The family of public service systems includes medical emergency system [11], fire-brigade deployment, public administration system design and many others. Mathematical models of the public service system design problem are often related to the p-median problem, where the p-median problem is formulated as a task of determination of at most p network nodes as facility locations. In real problems, the number of serviced customers takes the value of several thousands and the number of the possible facility locations can take this value as well. The number of possible service center locations impacts the computational time. To obtain good decision on the facility location in any serviced area, a mathematical model of

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 25–33

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the problem can be formulated and some of the mathematical programming methods can be applied to find the optimal solution. Reese summarized the exact solution methods for the p-median problem in [4]. Mladenovic summarized the heuristic methods in [8]. Avella, Sassano and Vasil´ev presented a branch-and-price-and-cut algorithm to solve large-scale instances of the p-median problem in [9]. Balinski provided an early integer programming formulation of the plant location problem that has historically been adapted to the p-median problem in [10]. Erlenkotter designed the approach for solving of the location problem in [1]. Erlenkotter approach is based on the theory of duality and the branch and bound algorithm. The suggested approach realizes an algorithm DualLoc. Körkel continued and improved the Erlenkotter approach and designed an algorithm PDLoc in [2]. Janáček and Buzna [3] improved the Erlenkotter and Körkel approach and designed an algorithm BBDual for solving the uncapacitated facility location problem. BBDual was extended by the Lagrangean relaxation of the constraint which limits number of the located center. Usage of the Lagrangean relaxation allowed to solve the p-median problem with the Erlenkotter approach. The Algorithm pMBBDual [5] was designed for solving the p-median problem with the Erlenkotter approach and the Lagrangean relaxation. If the algorithm pMBBDual does not provide the optimal solution, then we would like to improve the near-to-optimal solution by some heuristic.

2 Problem formulation

The p-median location problem finds the optimal location of exactly p facilities, so that the sum of the distances between customers and their closest facilities, measured along the shortest paths, is minimized. The location problem consists of a placing facility in some sites of a given finite set I such as hospitals, police stations, warehouses and the customers from a given finite set J such as people, patients in hospital or villages and cities. The costs of the optimal deployment of facilities in the specific network constitute the fixed charges f_i and the costs c_{ij}. The fixed charges f_i introduce costs for the facility location at the location i. The costs c_{ij} introduce costs for the demand satisfaction of a j-th customer from the location i. The formulated p-median location problem can be modeled using of the following notation. Let the decision of the service center location at the place $i \in I$ be modeled by a zero-one variable $y_i \in \{0,1\}$ which takes the value of 1, if the center is located at i, otherwise it takes the value of 0. In addition, the variables $z_{ij} \in \{0,1\}$ for each $i \in I$ and $j \in J$ are introduced to assign a customer j to a possible location i by the value of one. The maximal number of the facility locations introduces a constant p. The p-median location model can be formulated as follows:

Minimize
$$\sum_{i \in I} f_i y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} z_{ij} \tag{1}$$

Subject to:
$$\sum_{i \in I} z_{ij} = 1 \qquad \forall i \in I \tag{2}$$
$$z_{ij} \leq y_i \qquad \forall i \in I, \forall j \in J \tag{3}$$
$$\sum_{i \in I} y_i \leq p \tag{4}$$
$$y_i \in \{0,1\} \qquad \forall i \in I \tag{5}$$
$$z_{ij} \in \{0,1\} \qquad \forall i \in I, \forall j \in J \tag{6}$$

The objective function (1) minimizes the total costs of the p-median location problem which consists of the fixed charges f_i and the costs c_{ij}. The constraints (2) ensure that each customer is assigned to the exactly one possible service center location. Binding constraints (3) enable to assign a customer to a possible location i, only if the service center is located at this location. The constraint (4) bounds the number of the located service centers. The obligatory conditions in the mathematical model are (5) and (6). This location problem without the condition (4) gives the uncapacitated facility location problem (UFLP). If the location problem contains the condition (4) and the fixed charges f_i is equal zero for each $i \in I$ then it becomes p-median problem. Our p-median location problem is the combination of the UFLP and the p-median problem.

3 Solution method

Algorithm pMBBDual [5] provides us a possibility of solving the p-median problem with an iterative approach. The main advantage of the algorithm pMBBDual is the transformation of the p-median location problem to the UFLP. The mathematical model (1–6) using the Lagrangean relaxation of the constraint (4) which limits number of located centers is modified as follows:

Minimize
$$\sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} c_{ij} z_{ij} + lg(\sum_{i\in I} y_i - p) \tag{7}$$

Subject to:
$$\sum_{i\in I} z_{ij} = 1 \qquad \forall i \in I \tag{8}$$
$$z_{ij} \leq y_i \qquad \forall i \in I, \forall j \in J \tag{9}$$
$$y_i \in \{0,1\} \qquad \forall i \in I \tag{10}$$
$$z_{ij} \in \{0,1\} \qquad \forall i \in I, \forall j \in J \tag{11}$$

The solution of the mathematical model (7–11) represents one iteration of the algorithm pMBBDual. The quality and the feasibility of the solution of the suggested approach depends on a suitable setting of the Lagrangean multiplier lg. The suitable value of the multiplier can be obtained by a bisection algorithm. The comparison between XPRESS-IVE and the pMBBDual showed that the algorithm pMBBDual does not provide the optimal solution for the location problem all the time. We can obtain the optimal solution of the model (1–6) by repeating the solution of the model (7–11) with a change of the Lagrange multiplier lg until the last member of the objective function (7) is equal to zero. If a last member of the objective function (7) is non-equal to zero then we obtain some solution. The value of the obtained solution provides the lower bound (12) of the problem (1–6) which is written in a relation:

$$OF_{RP} = \sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} c_{ij} z_{ij} + lg(\sum_{i\in I} y_i - p) \tag{12}$$

The relation (13) provides the value of the obtained feasible solution of the problem (1–6):

$$OF_{NP} = \sum_{i\in I} f_i y_i + \sum_{i\in I}\sum_{j\in J} c_{ij} z_{ij} \tag{13}$$

We obtain some feasible solution, but a value of the optimal solution is between the values of the non-relaxation problem solving OF_{NP} (13) and the LP-relaxation solving OF_{RP} (12). If our approach does not obtain the optimal solution then we improve the near-to-optimal solution by the heuristic, which will be presented in the next chapter.

3.1 An exchange 1–1 heuristic with the reallocation of locations

An exchange 1–1 heuristic with the reallocation of locations works on the assumption that we have the best feasible and infeasible solution. The best feasible solution is the solution where number of locations is smaller and the nearest to p locations. The solution obtained by the algorithm pMBBDual is the best feasible solution. The best infeasible solution is the solution where number of locations is bigger and the nearest to p locations. The exchange 1–1 heuristic with the reallocation of locations has 3 phases:

Phase 1 – A reallocation of locations
A phase of the reallocation consists in creating of sets I_S and I_D from the locations in the best feasible and infeasible solution. The set I_S consists of the same locations in both solutions and the set I_D consists of the different locations in both solutions.

Phase 2 – An addition of locations to p
A phase of the addition consists in a separation of the set I_D to the two subsets I_{DI} and I_{DN}. The locations from I_{DI} are included in the set I_S and they create initial solution which will be improved. The set I_{DI} consists of the included locations. The set I_{DN} consists of the locations which are not included in the solution.

Phase 3 – A searching of the suitable exchange and its realization
A phase 3 searches the combination of locations which improves the actual solution. We exchange only one location from the set I_{DI} and the location from the set I_{DN}. If we obtain an exchange which improve the actual solution then we realize the exchange of the locations and update sets of I_{DI} and I_{DN}. The improving exchange can be obtained by the strategy first admissible or best admissible.

The exchange 1–1 heuristic ends when we do not find an improving exchange.

4 Numerical experiments

All numerical experiments mentioned in this paper were performed on a PC equipped with Intel(R) Core(TM) i7 Q720 1.6 GHz processor, 8 GB RAM. The tested benchmarks consist of the Slovak cities and villages in Slovak road network. The benchmarks in the Table 1 create the cost matrix consisted of only a distance between the places. The benchmarks in the Table 2 create the cost matrix consisted of a distance between the sites and the demands of the villages or the cities. We compare the solution obtained by our approach and the optimal solution obtained by the universal IP solver XPRESS-IVE in the computational time and the quality of solutions. The quality of solutions indicates the value of the objective function OF and the GAP_{ES} and the GAP_{LB}. The GAP_{ES} represents the difference between the values of the best found solution OF_{V1} and the exact OF_{ES} one expressed in the percentage of the exact solution OF_{ES} as follows:

$$GAP_{ES} = \frac{OF_{V1} - OF_{ES}}{OF_{ES}} * 100 \tag{14}$$

The GAP_{LB} corresponds to the difference between the value of the best found solution OF_{V1} and the lower bound OF_{RP} expressed in the percentage of the lower bound OF_{RP} as follows:

$$GAP_{LB} = \frac{OF_{V1} - OF_{RP}}{OF_{RP}} * 100 \tag{15}$$

4.1 Experiments 1

The experiments in the Table 1 was realized on the benchmarks from the set of all 315 places in the district of Žilina (Figure 1).

In the Table 1 a column p gives the maximal number of the facility locations, columns $t(s)$ give the computational time in seconds, columns NoF give number of the facility locations for the individual solution methods. A column OF_{ES} gives the value of the optimal solution, a column OF_{NP} gives the value of solving the non-relaxed problem (1–6), a column OF_{RP} gives the value of the solution LP-relaxation – the lower bound of the optimal solution, a column OF_{V1} gives the value of the final solution of the exchange 1–1 heuristic with the reallocation of locations. A column GAP_{EX} gives the representation in the percentage obtained by the relation (14) and a column GAP_{LB} gives the approximate representation in the percentage obtained by the relation (15). The time $t(s)$ in V1 gives the total time of the algorithm pMBBDual and the improving heuristic.

The experiments in the Table 1 shows that the algorithm pMBBDual gets better time in compared to the universal solver XPRESS-IVE, but the obtained solution by the algorithm pMBBDual can be optimal or near-to optimal. The obtained near-to optimal solution can be improved using the improving heuristic. Using of the improving heuristic to the obtained solution shows us the possibility of obtaining the optimal solution or better near-to-optimal solution at the expense of increasing the computational time. The computational time of the algorithm pMBBDual with the improving heuristic is better than the time obtained by the universal solver XPRESS-IVE. The GAP_{ES} shows that the difference in percentage is not worse than 1 %. Based on the experiments in the Table 1 the improving heuristic can provide the very near-to-optimal solution. The universal IP solvers are limited for solving the large problems. If we cannot obtain the value of the optimal solution with XPRESS-IVE, we use the lower bound OF_{RP} for the comparison of solutions. The GAP_{LB} is approximate representation in the percentage because the value of optimal solution do not need to equal the lower bound. The distortion is demonstrated in the Table 1 for the value p equals 210, where GAP_{LB} is 10,94 % and GAP_{ES} is only 0,47 %.

4.2 Experiments 2

Experiments in the Table 2 was realized on the benchmarks from the set of the customers consisting of all cities and villages and the set of the candidates consisting of the 1000 biggest villages and cities in Slovak Republic (Figure 2). In the Figure 2 red points give the candidates and all points give customers.

In the Table 2 a column p gives the maximal number of the facility locations, columns $t(s)$ give the computational time for the solution methods in seconds, a column NoF gives number of the facility locations obtained by the algorithm pMBBDual without the heuristic. A column OF_{NP} gives the value of solving the non-relaxation problem (1–6), a column OF_{RP} gives the value of the solution of the LP-relaxation – the lower bound of the optimal solution, a column OF_{V1} gives the value of the final solution of the exchange 1–1 heuristic with the reallocation of locations. A column GAP_{LB} gives the approximate representation in the percentage obtained by the relation (15). The time $t(s)$ in V1 gives total time of the algorithm pMBBDual and the improving heuristic.

The experiments in the Table 2 shows that we obtain the optimal solution by the algorithm pMBBDual with the improving heuristic for all selected value of p. The value of the optimal solution is equal to the lower bound of the solution. If we compare the computational times

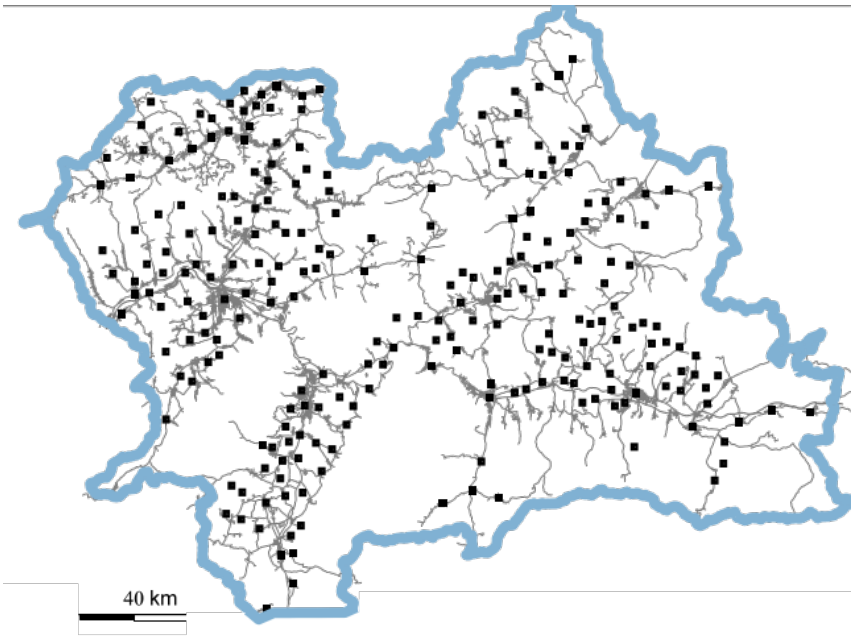


Figure 1 Deployment of cities and villages in district of ŽIlina.

Table 1 Comparison of solutions XPRESS-IVE and pMBBDual without and with heuristic.

	XPRESS-IVE			pMBBDUal				V1			
p	$t(s)$	OF_{ES}	NoF	$t(s)$	OF_{NP}	NoF	OF_{RP}	$t(s)$	OF_{V1}	GAP_{ES}	GAP_{LB}
15	14,13	2803	15	2,52	2803	15	2803	–	–	–	–
30	18,74	1832	30	1,78	1832	30	1832	–	–	–	–
45	13,65	1400	45	2,15	1420	44	1400	2,60	1401	0,07	0,07
60	16,92	1138	60	2,65	1183	57	1138	3,14	1138	0,00	0,00
75	13,6	944	75	2,11	977	72	944	2,20	945	0,11	0,11
90	13,53	801	90	1,88	828	87	801	1,91	802	0,12	0,12
105	13,69	686	105	1,64	707	102	686	2,70	690	0,58	0,58
120	13,69	595	120	2,01	637	113	595	4,31	598	0,50	0,50
135	13,5	516	135	1,88	637	113	505	2,82	517	0,19	2,38
150	13,59	441	150	1,85	506	137	441	4,26	444	0,68	0,68
165	13,72	372	165	1,98	506	137	366	4,04	374	0,54	2,19
180	14,12	312	180	1,85	340	173	312	13,32	313	0,32	0,32
195	14,45	257	195	1,88	340	173	252	10,17	257	0,00	1,98
210	13,96	212	210	1,85	340	173	192	5,86	213	0,47	10,94
225	14,4	167	225	1,88	206	212	167	3,00	167	0,00	0,00
240	14,2	123	240	1,61	206	212	122	3,04	123	0,00	0,82
255	14,29	93	255	1,78	107	248	93	2,00	93	0,00	0,00
270	13,77	63	270	1,74	107	248	63	1,98	63	0,00	0,00
285	13,71	33	285	1,58	107	248	33	2,01	33	0,00	0,00
300	13,74	14	300	1,58	23	291	14	1,85	14	0,00	0,00

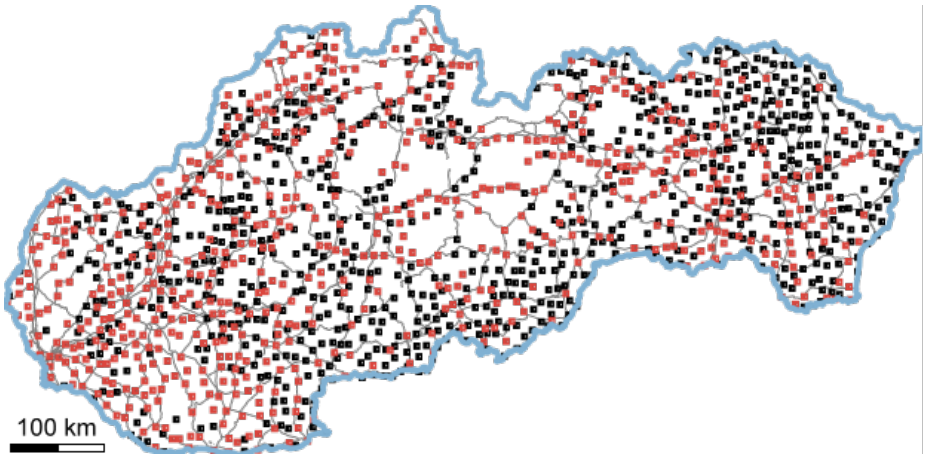


Figure 2 Deployment of cities and villages in Slovak Republic.

Table 2 Results of algorithm pMBBDual without and with the improving heuristic.

	pMBBDUal				V1		
p	$t(s)$	OF_{NP}	NoF	OF_{RP}	$t(s)$	OF_{V1}	$GAP_{LB}(\%)$
50	100,8	535448	50	535448	–	–	–
100	35,4	321869	100	321869	–	–	–
150	65,7	248373	150	248373	–	–	–
200	113,0	207011	200	207011	–	–	–
250	104,5	178645	250	178645	–	–	–
300	132,4	160076	297	159044	143,3	159044	0,00
350	68,2	144076	350	144076	–	–	–
400	68,4	132427	400	132427	–	–	–
450	54,9	123775	449	123615	65,3	123615	0,00
500	71,9	116650	498	116388	72,6	116388	0,00
550	52,3	110558	549	110450	52,7	110450	0,00
600	46,0	105418	600	105418	–	–	–
650	46,1	101209	650	101209	–	–	–
700	41,0	98229	694	97869	41,9	97869	0,00
750	38,6	95308	747	95158	39,4	95158	0,00
800	39,0	93000	796	92832	40,3	92832	0,00
850	38,7	90976	848	90906	40,6	90906	0,00
900	39,1	89648	889	89340	42,1	89340	0,00
950	39,4	88264	945	88164	41,3	88164	0,00
1000	0,3	87427	1000	87427	–	–	–

of the algorithm pMBBDual without and with the improving heuristic then the time of the realization of the improving heuristic is a few seconds.

5 Conclusions

Solving the p-median location problem in the public service system design is NP-hard problem. The optimal solution of the problem can be obtained by the universal IP solvers only for smaller instances of the problem. The universal IP solvers are very time-consuming and often fail when a large instance is solved. Our approach to the problem was based on the Erlenkotter procedure for solving of the uncapacitated facility location problem and on the Lagrangean relaxation of the constraint which limits number of the located center. We designed algorithm pMBBDual which does not provide optimal solution for location problem every time. So we tried to improve the obtained near-to-optimal solution with some heuristic. We designed the exchange 1–1 heuristic with the reallocation of locations. The resulting solution of our approach with the exchange heuristic and the optimal solution obtained by the universal IP solver XPRESS-IVE were compared in the computational time and the quality of solutions. Based on the numerical experiments we review that the solution obtained by the algorithm pMBBDual is possible to improve. We cannot obtain the optimal solution from near-to-optimal solution with the suggested improving heuristic every time. But the improving heuristic can provide the very near-to-optimal solution in many instances of the solved problem.

We improved the obtained solution by one heuristic, but in the future we would like to design the other improving heuristics and choose the best heuristic to our approach. We would like to generalize the Erlenkotter approach, design an algorithm with the Erlenkotter approach for solving the p-median location problem which is not iterative and compare algorithm pMBBDual with the improving heuristic, the algorithm with Erlenkotter approach which is not iterative and the Z-Erlange and branch algorithm [12] in the computational time and the quality of the obtained solution for the large-scale problems.

Acknowledgements. This work was supported by the research grants VEGA 1/0296/12 "Public Service Systems with Fair Access to Service" and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks".

References

1 Erlenkotter, D.: *A Dual-Based Procedure for Uncapacitated Facility Location*. Operations Research, Vol. 26, No. 6, November-December 1978, pp. 992–1009

2 Körkel, M.: *On the exact solution of large – scale simple plant location problem*. In European Journal of Operational Research 39, North Holland, pp. 157–173, 1989.

3 Janáček, J. and Buzna, L.: *An acceleration of Erlenkotter-Körkels algorithms for the uncapacitated facility location problem*. Annals of Operations Research, vol. 164, 97–109, 2008.

4 Reese, J.: *Solution methods for the p-median problem*. Networks, vol. 48, no. 3, pp. 125–142, 2006.

5 Bendík, J.: *Exact algorithm for Location Problem Solving in Public Service System Design*. Úlohy diskrétní optimalizace v dopravní praxi 2013,Pardubice 28.-29.10.2013, pp. 7–15,ISBN 978-80-7395-744-5.

6 Garey, M. R. and Johnson, D. .S.: *Computers and intractibility: A guide to the theory of NP-completeness* W.H. Freeman and Co., San Francisco, 1979.

7 Kariv, C. and Hakimi, S. L.: *An algorithmic approach to network location problems*. II. The p-medians, SIAM J Appl Math 37, pp. 539–560, 1979.

8 Mladenovic, N. and Brimberg, J. and Hansen, P. and Moreno-Pereéz, J. A.: *The p-median problem: a survey of metaheuristic approaches*. Eur J Oper Res 179, pp. 927–939, 2007.

9 Avella, P. and Sassano, A. and Vasil'ev, I.: *Computational study of large-scalep-median problems*. Technical report 08-03, Università di Roma "La Sapienza," 2003.

10 Balinski, M.: *Integer programming, methods, uses and computation*. Manage Sci 12), pp. 253–313, 1965.

11 Marianov, V. and Serra, D.: *Location problems in the public sector*. In: Drezner, Z. (ed.) et al. Facility location. Applications and theory. Berlin: Springer, pp 119–150, 2002.

12 Garía, S. and Labbé, M. and Marín, A.: *Solving large p-median problems with a radius formulation*. IN-FORMS Journal on Computing 23 (4) 546–556, 2011.

A new approach to modelling nonlinear time series: Introducing the ExpAR-ARCH and ExpAR-GARCH models and applications

Paraskevi Katsiampa

Loughborough University
Loughborough, England
P.Katsiampa@lboro.ac.uk

Abstract

The analysis of time series has long been the subject of interest in different fields. For decades time series were analysed with linear models. Nevertheless, an issue that has been raised is whether there exist other models that can explain and fit real data better than linear ones. In this paper, new nonlinear time series models are proposed (namely the ExpAR-ARCH and the ExpAR-GARCH), which are combinations of a nonlinear model in the conditional mean and a nonlinear model in the conditional variance and have the potential of explaining observed data in various fields. Simulated data of these models are presented, while different algorithms (the Nelder-Mead simplex direct search method, the Quasi-Newton line search algorithm, the Active-Set algorithm, the Sequential Quadratic Programming algorithm, the Interior Point algorithm and a Genetic Algorithm) are used and compared in order to check their estimation performance when it comes to these suggested nonlinear models. Moreover, an application to the Dow Jones data is considered, showing that the new models can explain real data better than the AR-ARCH and AR-GARCH models.

1998 ACM Subject Classification G.3 Probability and Statistics, G.1.6 Optimization

Keywords and phrases Nonlinear time series, ExpAR-ARCH model, ExpAR-GARCH model

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.34

1 Introduction

During the last century considerable achievements have been made in both theoretical and empirical linear time series analysis. The Autoregressive (AR) model of Yule (1927) and the Autoregressive Moving Average (ARMA) model of Box and Jenkins (1970) are the two most noticeable examples of linear models which have found many applications in real life data. Linear models have many advantages, such as good fitting and predictive ability, which is the main reason why they have been used so much. However, there are time series that exhibit nonlinear characteristics, in which case, linear time series models can be too restrictive and if our aim is a more profound analysis of how series are generated, we need to allow for more general models.

Hence, the limitations of linear models have raised the issue of whether there exist other models that can explain and predict better time series with such characteristics. This issue resulted in the expansion of the linear models in the literature and in the development of various nonlinear models, e.g. nonlinear models in conditional mean and nonlinear models in conditional variance, all of which have attempted to explain and forecast more accurately specific time series.

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 34–51

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The classic nonlinear models in conditional mean are the Threshold Autoregressive (TAR) model of Tong (1977) (along with its special cases, the Self-Exciting Threshold Autoregressive (SETAR) model (Tong and Lim (1980)) and the Smooth Transition Autoregression (STAR) model (Chan and Tong (1986))), the Exponential Autoregressive (ExpAR) model of Ozaki (1980) and the Bilinear model of Granger and Andersen (1978). On the other hand, the most characteristic examples of nonlinear models in conditional variance are the Autoregressive Conditional Heteroskedasticity (ARCH) and the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models, introduced by Engle (1982) and Bollerslev (1986) respectively.

Since the introduction of these models, there have been introduced many nonlinear models, some trying to explain nonlinearities in the conditional mean and others attempting to interpret nonlinearities in the conditional variance. Nevertheless, there are some time series exhibiting asymmetries which could be better explained by models that have both a nonlinear conditional mean and a changing conditional variance, but there has not been much work on combining these two forms of nonlinearity. Tong (1990), p. 116, was the first to suggest combining the first-generation models in order to produce second-generation models, as he called them, giving as examples the specification of a SETAR-ARCH model and a Bilinear-ARCH model, which combine a SETAR or a Bilinear model, respectively, for the conditional mean with a conditional variance following an ARCH model. Since then, such models have gradually become popular and are being used more widely, mainly in applications to financial data.

The class of these second-generation models that has been applied most though is the TAR-GARCH family, and especially the STAR-GARCH and STAR-STGARCH models. Applications of TAR-GARCH-type and SETAR-GARCH-type models can be found in Li and Lam (1995), in Li and Li (1996), in Amendola and Niglio (2000), in Osinska and Witkowski (2004), in Chiang and Doong (2001), and in Munoz, Marquez and Acosta (2007), while applications of STAR-GARCH-type models can be found in Lee and Li (1998), in Lundbergh and Terasvirta (1999), and in Busetti and Manera (2003).

A different class of models that combines a nonlinear conditional mean and conditional variance is the Exponential Autoregressive model with GARCH errors, which, however, has not been much used or developed. The first model of this class, introduced by LeBaron (1992) with the purpose of exploring the relationship between volatility and serial correlation for different stock return series at daily and weekly frequencies, was a combination of Bollerslev's (1986) GARCH model, Ozaki's (1980) ExpAR model and Stock's (1988) time deformation model. Later, Koutmos (1997) used an Exponential Autoregressive model for the conditional mean with a Threshold GARCH model for the conditional standard deviation (EAR-TGARCH) along with a Generalised Error Distribution, which was a generalised version of LeBaron's (1992) model, in order to study the daily stock returns in some equity markets of the Pacific Basin area and to examine if the behaviour of these markets are similar to the behaviour of developed ones.

In this paper, we suggest the ExpAR-ARCH and ExpAR-GARCH models, which are combinations of the pure ExpAR model of Ozaki (1980) for the conditional mean and the ARCH or GARCH model respectively for the conditional variance, and which have the potential of explaining and forecasting nonlinear time series of various fields. It should be highlighted that these models are different from the ones proposed by LeBaron (1992) and by Koutmos (1997) in the variable contained in the exponential term of the conditional mean model. Our models are in accordance with the ExpAR model suggested by Ozaki (1980) containing the lag of the variable in the exponential term, while LeBaron's (1992) model and

Koutmos' (1997) model contain the conditional variance instead.

The paper is organised as follows: In section 2, the models are introduced. In section 3, the estimation method is presented. Some simulation results are shown in section 4, while in section 5 an application to real data is considered. Finally, some concluding remarks are made in section 6.

2 Models

In this section, the suggested models are presented. These consist of a nonlinear model for the conditional mean (ExpAR) and a nonlinear model for the conditional variance (ARCH or GARCH).

Let y_t be a time series generated by a stationary process. The Exponential Autoregressive model of order s with heteroscedastic errors is defined as:

$$y_t = c + \sum_{i=1}^{s} \{\phi_i + \pi_i \cdot exp(-\gamma \cdot y_{t-1}^2)\} \cdot y_{t-i} + u_t, \tag{1}$$

where

$$u_t = \epsilon_t \cdot \sqrt{h_t}, \tag{2}$$

$$\epsilon_t \sim n.i.d.(0,1), \text{ and} \tag{3}$$

$$h_t = h(n) = n' \cdot z_t. \tag{4}$$

In the case of the ExpAR(s)-ARCH(q) model, we have

$$n = (\alpha_0, \alpha_1, \ldots, \alpha_q)', \tag{5}$$

$$z_t = (1, u_{t-1}^2, \ldots, u_{t-q}^2), \tag{6}$$

$$\alpha_0 > 0 \text{ and } \alpha_i \geq 0, i > 0, i = 1, \ldots, q, \tag{7}$$

while in the case of the ExpAR(s)-GARCH(p,q) model, we have

$$n = (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)', \tag{8}$$

$$z_t = (1, u_{t-1}^2, \ldots, u_{t-q}^2, h_{t-1}, \ldots, h_{t-p}), \tag{9}$$

$$\alpha_0 > 0, \alpha_i \geq 0, i > 0, i = 1, \ldots, q, \text{and } \beta_j \geq 0, j > 0, j = 1, \ldots, p. \tag{10}$$

3 Estimation

The estimation procedure that is used in this research is maximum likelihood. Assuming that the sequence u_t is identically normal distributed and conditioning on the observation at time $t = 0$, y_0, the conditional log-likelihood function is

$$L_T(\theta) = \sum_{i=1}^{T} l_t(\theta), \tag{11}$$

where

$$l_t = -\frac{1}{2} \cdot \log 2\pi - \frac{1}{2} \cdot \log h_t - \frac{1}{2} \cdot \frac{u_t^2}{h_t}, \tag{12}$$

is the log-likelihood at time t, which means that the overall conditional log-likelihood function is

$$L_t(\theta) = -\frac{T}{2} \cdot \log 2\pi - \frac{1}{2} \cdot \sum_{i=1}^{T} \log h_t - \frac{1}{2} \cdot \sum_{i=1}^{T} \frac{u_t^2}{h_t}. \tag{13}$$

The gradient of the overall conditional log-likelihood is defined as

$$G_T = [\partial L_T/\partial b', \partial L_T/\partial \omega'], \tag{14}$$

while the gradient of the log-likelihood function at time t is given by

$$g_T = [\partial l_t/\partial b', \partial l_t/\partial \omega'], \tag{15}$$

where b is the vector of conditional mean parameters and ω is the vector of the conditional variance parameters.

In the previous formulae of the estimation part, u_t should be replaced by

$$y_t - c - \sum_{i=1}^{s} \{\phi_i + \pi_i \cdot exp(-\gamma \cdot y_{t-1}^2)\} \cdot y_{t-i}. \tag{16}$$

However, maintaining u_t makes the notation less complicated and keeps it close to Bollerslev (1986).

4 Simulations

4.1 Methodology

In this section simulated series of the proposed models are presented. Since new models are introduced, it is important to simulate them. In this way we can see what characteristics real data, which could be described by them, would have. It should be emphasised that here only the first order models are considered, as it is well-known that low orders of nonlinear models can capture the biggest part of nonlinearity.

The simulated series are compared with the well-known AR-ARCH and AR-GARCH models, which can be considered as benchmarks. The comparison is made by using the same

Table 1 Moments of the simulated series.

	ExpAR(1)-ARCH(1)	ExpAR(1)-GARCH(1,1)	AR(1)-ARCH(1)	AR(1)-GARCH(1,1)
Mean	0.5381	0.4838	0.4213	0.4179
Standard deviation	0.3045	0.7523	0.3113	0.7636
Skewness	-0.2438	-0.2637	0.0222	-0.0511
Kurtosis	42.916	83.326	43.487	81.298

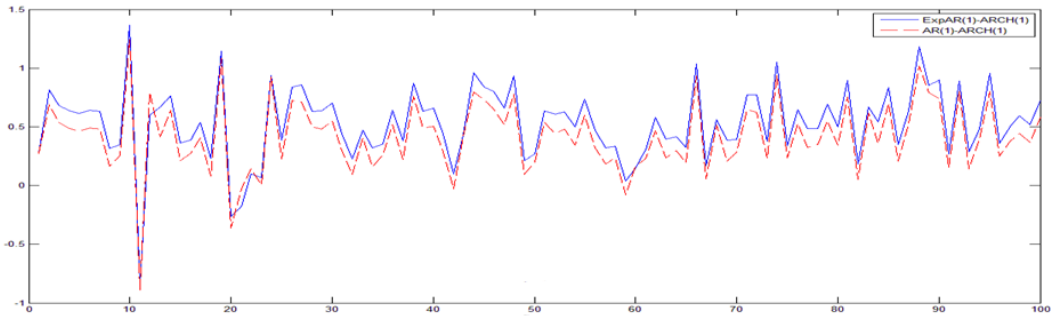


Figure 1 Simulated series: ExpAR(1)-ARCH(1) and AR(1)-ARCH(1).

values for the common parameters (same c, ϕ_1, α_0 and α_1 for all the models, same β_1 for the models with GARCH errors, and same π_1 and γ for the models that are described by an ExpAR model for the conditional mean) and the same number of simulated data ($T = 3400$).

The parameter values were chosen as follows: In order to be consistent with the conditional variance model restrictions, we chose non-negative conditional variance parameters values. We also wanted a stationary conditional mean model and a stationary conditional variance model. Hence, we set the ϕ_1 parameter to a value which is smaller than one in absolute value, and we set the α_1 parameter and the sum of the α_1 and β_1 parameters to be smaller than one in the case of ARCH and GARCH errors respectively.

Moreover, since estimation plays a very important role when fitting data to models, it is essential to estimate the coefficients of the models in order to see how close the assumed and estimated parameters are and to suggest effective estimating methods. Furthermore, various algorithms have been used and compared in order to check their estimation performance when it comes to the suggested nonlinear models and to the classic AR-ARCH and AR-GARCH models. More specifically, the algorithms used are the Nelder-Mead simplex direct search method (NM), the Quasi-Newton line search algorithm (QN), the Active-Set algorithm (AS), the Sequential Quadratic Programming algorithm (SQP), the Interior Point algorithm (IP) and a Genetic Algorithm (GA).

The NM and the QN methods solve unconstrained optimisation problems, while the GA, the AS, the SQP and the IP algorithms solve the constrained optimisation problem. The GA was used to solve the unconstrained optimisation problem as well, but it never gave good results, as it gave some negative estimates for the ARCH or GARCH parameters, resulting in complex values of the log-likelihood function and of the standard errors, and therefore its results are not displayed.

In Table 1 there can be found the moments of each simulated series, while Figures 1-6 show the simulated series and their autocorrelation and partial autocorrelation plots. Table 2 shows the arbitrary initial values used when running the algorithms. The values of the real parameters for the simulations are displayed in the first column of Tables 3-6, while in the remaining columns the results obtained from the algorithms are reported, including the

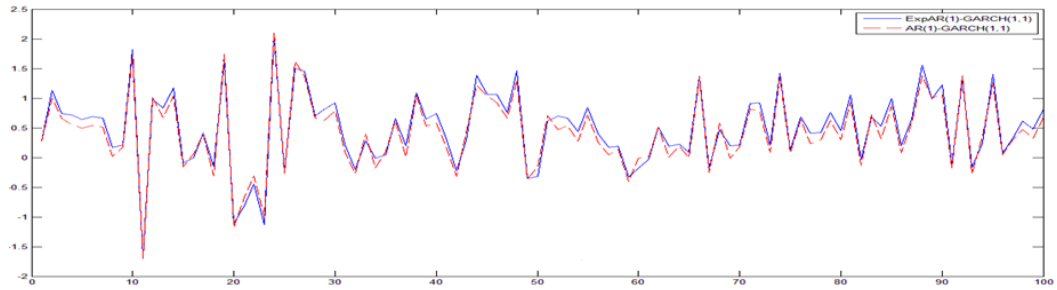


Figure 2 Simulated series: ExpAR(1)-GARCH(1,1) and AR(1)-GARCH(1,1).

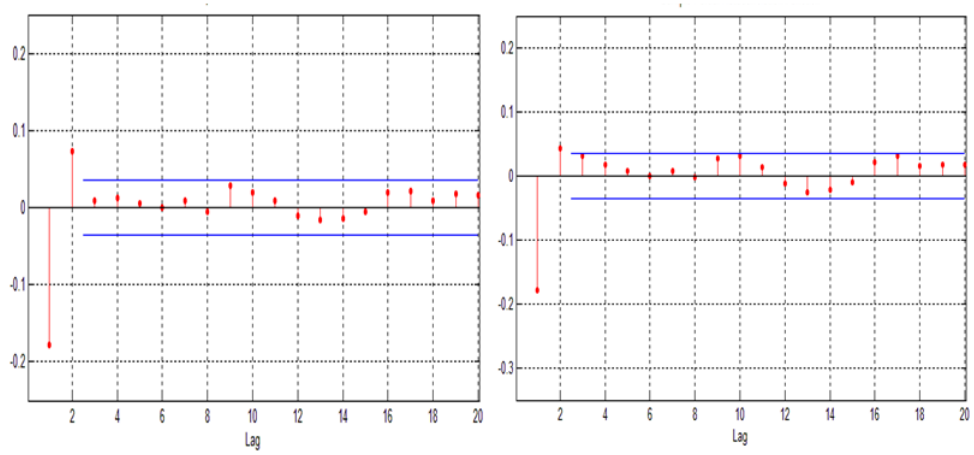


Figure 3 Autocorrelation (a) and Partial Autocorrelation (b) plot of ExpAR(1)-ARCH(1).

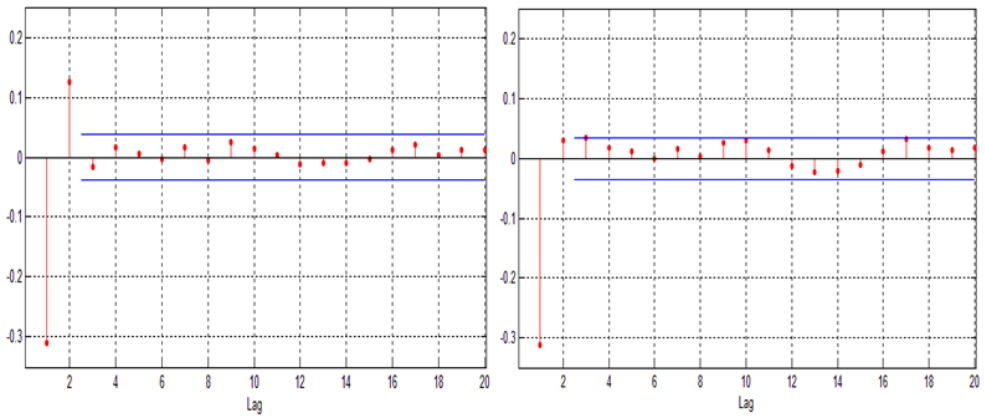


Figure 4 Autocorrelation (a) and Partial Autocorrelation (b) plot of AR(1)-ARCH(1).

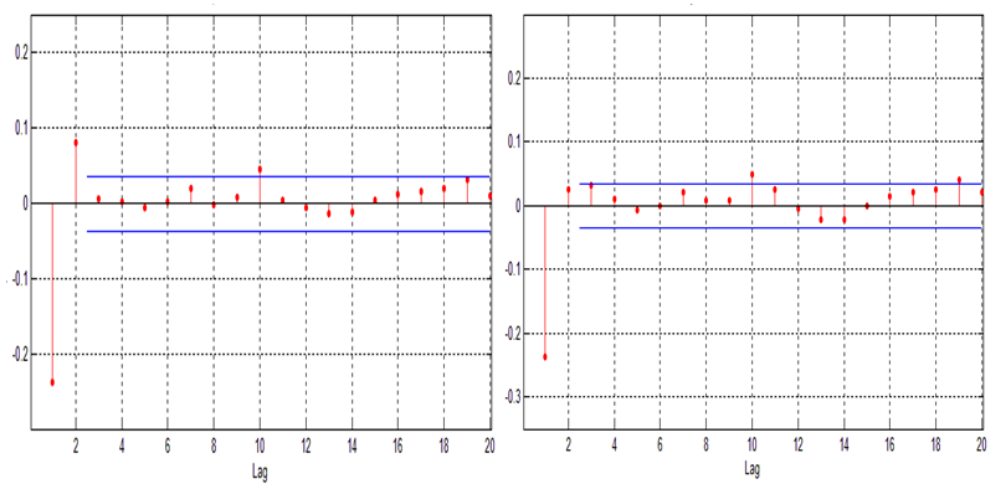


Figure 5 Autocorrelation (a) and Partial Autocorrelation (b) plot of ExpAR(1)-GARCH(1,1).

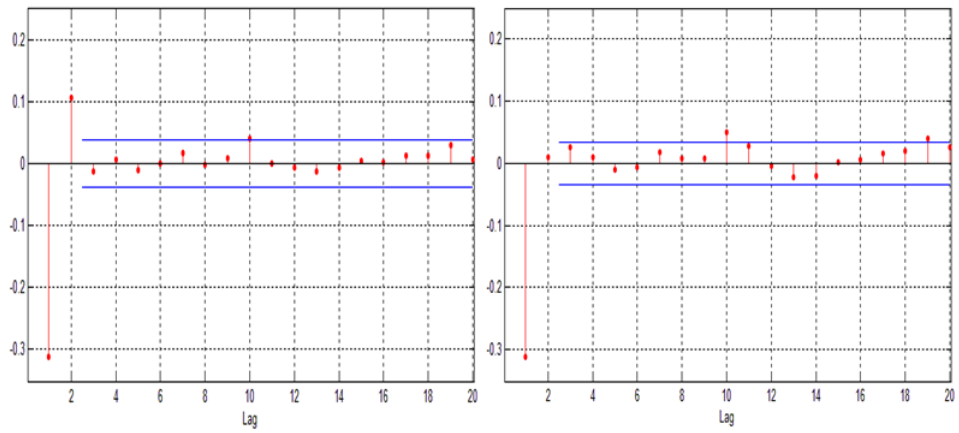


Figure 6 Autocorrelation (a) and Partial Autocorrelation (b) plot of AR(1)-GARCH(1,1).

estimates for the parameters, the respective standard errors, the value of the log-likelihood function, and the number of iterations/generations required for the convergence of the algorithm to the optimum. In addition, once we obtain the estimates from the algorithms, the algorithms are run again, but then the initial values are not arbitrary. Instead, they are chosen accordingly to which algorithm gave better results in terms of closeness to the real values, so that the accuracy and speed of the algorithms are tested when the initial guess for the values of the parameters is indeed close to the real parameters.

4.2 Results

4.2.1 Using arbitrary initial values for the algorithms

According to the results obtained, we notice that for the ExpAR(1)-ARCH(1) model only the QN algorithm gave estimates which are close to all the true parameter values. Moreover, the QN algorithm required the lowest number of iterations in order to converge to a solution and gave the highest log-likelihood value. On the other hand, the AS algorithm failed and didn't give any results, while the remaining algorithms gave relatively good estimates only for the constant c and the parameters of the ARCH(1) model, α_0 and α_1. Moreover, the SQP and IP

Table 2 Initial values of the algorithms.

	ExpAR(1)-ARCH(1)	**ExpAR(1)-GARCH(1,1)**	**AR(1)-ARCH(1)**	**AR(1)-GARCH(1,1)**
c	0.0010	0.0010	0.0010	0.0010
ϕ_1	0.6000	0.6000	0.6000	0.6000
π_1	0.3000	0.3000	-	-
γ	1	1	-	-
α_0	0.0010	0.0010	0.0010	0.0010
α_1	0.1000	0.1000	0.1000	0.1000
β_1	-	0.2000	-	0.2000

Table 3 Estimation results for ExpAR(1)-ARCH(1).

	ExpAR(1)-ARCH(1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5958 (0.0094)	0.5689 (0.0112)	NaN (NaN)	0.5957 (0.0094)	0.5957 (0.0094)	0.6463 (0.0102)
$\phi_1 = -0.30$	18.3029 (0+9.7672i)	-0.3366 (0.0732)	NaN (NaN)	18.8319 -20.0583	19.3354 -9.3605	-0.7759 (0.1661)
$\pi_1 = 0.50$	-18.2896 (0+9.7662i)	0.4922 (0.0558)	NaN (NaN)	-18.8185 -20.0555	-19.3219 -9.3596	0.6571 (0.1729)
$\gamma = 1.20$	-0.0109 (0+0.0057i)	1.1357 (0.3388)	NaN (NaN)	-0.0106 (0.0111)	-0.0103 (0.0050)	0.1857 (0.0543)
$\alpha_0 = 0.05$	0.0502 (0.0013)	0.0505 (0.0013)	NaN (NaN)	0.0502 (0.0013)	0.0502 (0.0013)	0.0546 (0.0015)
$\alpha_1 = 0.40$	0.4463 (0.0234)	0.4398 (0.0235)	NaN (NaN)	0.4463 (0.0234)	0.4463 (0.0234)	0.3675 (0.0202)
Log-likelihood	2.1720e+03	2.1799e+03	NaN	2.1720e+03	2.1720e+03	2.1461e+03
Iterations/ Generations	1018	69	400	170	115	112

methods gave the same estimates and standard errors for the parameters c, α_0 and α_1, while the estimates obtained by these two algorithms for the remaining parameters are close to each other, although very far away from the true values. Furthermore, the estimates obtained by the NM method are close to the estimates obtained by the SQP and IP algorithms for all the parameters. The GA gave good estimates only for the ARCH parameters.

In the case of the ExpAR(1)-GARCH(1, 1) model, the QN algorithm again performed quite well. In this model, however, we observe that the IP algorithm gave estimates that are close to the true values of the parameters as well. In fact, here the highest log-likelihood value was obtained by the IP algorithm and the second highest log-likelihood value was given by the QN algorithm, although the latter required slightly fewer iterations. Yet the difference between these two log-likelihood values is rather unimportant. In addition, the Genetic Algorithm under constrained optimisation gave good estimates for the parameters of the conditional mean model, although it overestimated the parameter γ. However, it didn't give very good estimates for the GARCH parameters, in contrast to the QN, IP and SQP methods. The AS algorithm failed here as well.

Table 4 Estimation results for ExpAR(1)-GARCH(1,1).

	ExpAR(1)-GARCH(1,1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.2307 (0.0091)	0.5733 (0.0101)	NaN (NaN)	0.6248 (0.0093)	0.5740 (0.0101)	0.5844 (0.0094)
$\phi_1 = -0.30$	6.2888 (-1.9683)	-0.3454 (0.0258)	NaN (NaN)	26.7355 (-10.5873)	-0.3432 (0.0255)	-0.3011 (0+0.0228i)
$\pi_1 = 0.50$	-6.0639 (-1.9649)	0.4714 (0.0384)	NaN (NaN)	-26.9161 (-10.5871)	0.4659 (0.0384)	0.5108 (0.0444)
$\gamma = 1.20$	-0.0072 (0.0021)	1.0593 (0.1730)	NaN (NaN)	-0.0007 (0.0013)	1.0652 (0.1738)	2.0012 (0+0.5525i)
$\alpha_0 = 0.05$	0.5076 (0+0.0279i)	0.0535 (0.0039)	NaN (NaN)	0.0561 (0.0041)	0.0536 (0.0039)	0.1210 (0.0098)
$\alpha_1 = 0.40$	0.3328 (0.0090)	0.4497 (0.0219)	NaN (NaN)	0.4612 (0.0225)	0.4493 (0.0219)	0.5254 (0.0239)
$\beta_1 = 0.50$	-0.0243 (0+0.0105i)	0.4713 (0.0194)	NaN (NaN)	0.4593 (0.0197)	0.4713 (0.0194)	0.2755 (0.0240)
Log-likelihood	-4.3460e+03	-2.94089e+03	NaN	-3.0024e+03	-2.94088e+03	-3.0296e+03
Iterations/ Generations	1417	74	400	109	80	116

Table 5 Estimation results for AR(1)-ARCH(1).

	AR(1)-ARCH(1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5628 (0.0059)	0.5631 (0.0059)	0.5627 (0.0059)	0.5627 (0.0059)	0.5627 (0.0059)	0.5660 (0.0060)
$\phi_1 = -0.30$	-0.3286 (0.0121)	-0.3298 (0.0121)	-0.3285 (0.0121)	-0.3285 (0.0121)	-0.3285 (0.0121)	-0.3321 (0.0122)
$\alpha_0 = 0.05$	0.0505 (0.0013)	0.0506 (0.0013)	0.0505 (0.0013)	0.0505 (0.0013)	0.0505 (0.0013)	0.0498 (0.0014)
$\alpha_1 = 0.40$	0.4390 (0.0234)	0.4398 (0.0235)	0.4390 (0.0234)	0.4390 (0.0234)	0.4390 (0.0234)	0.4912 (0.0264)
Log-likelihood	2.1800e+03	2.1800e+03	2.1800e+03	2.1800e+03	2.1800e+03	2.1773e+03
Iterations/ Generations	250	51	71	49	58	97

Table 6 Estimation results for AR(1)-GARCH(1,1).

	AR(1)-GARCH(1,1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.3062	0.5661	0.5660	0.5660	0.5660	0.5951
	(1.1215e-05-1.2259e-12i)	(0.0077)	(0.0077)	(0.0077)	(0.0077)	(0.0078)
$\phi_1 = -0.30$	0.2855	-0.3360	-0.3359	-0.3359	-0.3359	-0.3771
	(1.3589e-05+2.7361e-12i)	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0120)
$\alpha_0 = 0.05$	-0.0730	0.0533	0.0534	0.0534	0.0534	0.0626
	(1.0032e-09+1.3493e-15i)	(0.0039)	(0.0039)	(0.0039)	(0.0039)	(0.0048)
$\alpha_1 = 0.40$	11.451	0.4486	0.4483	0.4483	0.4483	0.4706
	(1.2296e-07 + 5.7040e-12i)	(0.0218)	(0.0218)	(0.0218)	(0.0218)	(0.0237)
$\beta_1 = 0.50$	0.5182	0.4727	0.4726	0.4727	0.4727	0.4428
	(2.1610e-12+4.8252e-06i)	(0.0193)	(0.0193)	(0.0193)	(0.0193)	(0.0211)
Log-likelihood	-2.8144e-06-9.9903e+02i	-2.9413e+03	-2.9413e+03	-2.9413e+03	-2.9413e+03	-2.9520e+03
Iterations/ Generations	558	57	87	58	37	73

When running the algorithms for estimating the AR(1)-ARCH(1) model, we notice that all the methods performed similarly and well. In fact, the IP, SQP and AS algorithms gave exactly the same estimates and standard errors, while the NM gave slightly different estimates for the constant c and the autoregressive parameter ϕ_1. The QN method and the GA performed similarly, but it could be said that the differences are rather unimportant, so that any of these algorithms could be used to estimate an AR(1)-ARCH(1) model. However, the algorithm that required the lowest number of iterations is the SQP (49) and then the QN (51).

Similar to the AR(1)-ARCH(1) model, when estimating the AR(1)-GARCH$(1,1)$ the IP, SQP and AS algorithms gave exactly the same estimates, while the QN method gave slightly different estimates, but again the differences are rather unimportant. All these four algorithms performed well. However, here the GA gave estimates that deviate more from the true values and the NM method did not perform well at all, even giving a negative value for the α_0 parameter. Here, the IP algorithm required the minimum number of iterations (37), while the QN and the SQP methods followed, requiring, 57 and 58 iterations respectively.

It is easily noticed that when estimating the above models, the QN algorithm, although solving the unconstrained problem, not only gave estimates close to the real values of the parameters in every single case, compared to the IP, SQP, AS, NM and GA methods, but also overall required a small number of iterations in order to converge to the solution. Hence, since the QN method performed well in every case, it seems logical to use the estimates obtained from it as initial guesses and run the algorithms again in order to check if there is any improvement in their performance, when we know that the initial values we give to the algorithms are indeed close to the true values. The initial values used for the second round of optimisations are shown in Table 7, while the respective estimation results can be found in Tables 8-11.

4.2.2 Using non-arbitrary initial values for the algorithms

When the initial guesses were closer to the true values of the parameters, we notice that for both the ExpAR(1)-ARCH(1) and the ExpAR(1)-GARCH$(1,1)$ models all the algorithms, apart from the GA which does not require initial values, performed quite well and in fact they gave the same estimates, while requiring a smaller number of iterations to reach a

Table 7 Initial values for the algorithms on the second round.

	ExpAR(1)-ARCH(1)	**ExpAR(1)-GARCH(1,1)**	**AR(1)-ARCH(1)**	**AR(1)-GARCH(1,1)**
c	0.5689	0.5733	0.5631	0.5661
ϕ_1	-0.3366	-0.3454	-0.3298	-0.3360
π_1	0.4922	0.4714	-	-
γ	11.357	10.593	-	-
α_0	0.0505	0.0535	0.0506	0.0533
α_1	0.4398	0.4497	0.4398	0.4486
β_1	-	0.4713	-	0.4727

Table 8 Estimation results for ExpAR(1)-ARCH(1) on the second round.

	ExpAR(1)-ARCH(1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5721 (0.0112)	0.5721 (0.0112)	0.5721 (0.0112)	0.5721 (0.0112)	0.5721 (0.0112)	0.8084 (0.0041)
$\boldsymbol{\phi_1 = -0.30}$	-0.3749 (0.0812)	-0.3749 (0.0812)	-0.3748 (0.0811)	-0.3748 (0.0812)	-0.3749 (0.0812)	-0.3334 (0+0.0060i)
$\boldsymbol{\pi_1 = 0.50}$	0.5113 (0.0642)	0.5113 (0.0642)	0.5112 (0.0641)	0.5113 (0.0642)	0.5113 (0.0642)	-0.3054 (0+0.0377i)
$\boldsymbol{\gamma = 1.20}$	0.9700 (0.2915)	0.9700 (0.2915)	0.9701 (0.2912)	0.9701 (0.2917)	0.9700 (0.2914)	26.610 (0+0.3613i)
$\boldsymbol{\alpha_0 = 0.05}$	0.0505 (0.0013)	0.0505 (0.0013)	0.0505 (0.0013)	0.0505 (0.0013)	0.0505 (0.0013)	0.0553 (0.0015)
$\boldsymbol{\alpha_1 = 0.40}$	0.4399 (0.0234)	0.4399 (0.0234)	0.4399 (0.0234)	0.4399 (0.0234)	0.4399 (0.0234)	0.4605 (0.0243)
Log-likelihood	2.1801e+03	2.1801e+03	2.1801e+03	2.1801e+03	2.1801e+03	1.9529e+03
Iterations/ Generations	212	33	12	13	31	113

solution, as would be expected. Moreover, now the AS method not only did not fail, but required the minimum number of iterations for both models as well.

Nevertheless, the solutions obtained now are slightly worse than the ones obtained before from the QN algorithm, as now the estimates for the parameters of the conditional mean model, and especially the estimate for the γ parameter in the case of the ExpAR(1)-ARCH(1) model, deviate slightly more from the true values, although the values of the log-likelihood are somewhat higher than before for both models. In addition, in contrast to the ExpAR(1)-ARCH(1) model, the estimate for the γ parameter for the ExpAR(1)-GARCH(1, 1) model was slightly improved.

Furthermore, for the ExpAR(1)-ARCH(1) model the GA here gave good estimates only for the ϕ_1 parameter and for the parameters of the conditional variance model, while for the ExpAR(1)-GARCH(1, 1) model it gave better estimates for the c, γ, α_0, α_1 and β_1 parameters, but worse estimates for ϕ_1 and π_1.

Table 9 Estimation results for ExpAR(1)-GARCH(1,1) on the second round.

	ExpAR(1)-GARCH(1,1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5740	0.5740	0.5740	0.5740	0.5740	0.5656
	(0.0101)	(0.0101)	(0.0101)	(0.0101)	(0.0101)	(0.0108)
$\phi_1 = -0.30$	-0.3432	-0.3432	-0.3432	-0.3432	-0.3432	-0.2130
	(0.0255)	(0.0255)	(0.0255)	(0.0255)	(0.0255)	(0.0167)
$\pi_1 = 0.50$	0.4659	0.4659	0.4659	0.4659	0.4659	0.2653
	(0.0384)	(0.0384)	(0.0384)	(0.0384)	(0.0384)	(0.0478)
$\gamma = 1.20$	10.652	10.652	10.652	10.653	10.652	11.389
	(0.1738)	(0.1738)	(0.1738)	(0.1738)	(0.1738)	(0.1594)
$\alpha_0 = 0.05$	0.0536	0.0536	0.0536	0.0536	0.0536	0.0602
	(0.0039)	(0.0039)	(0.0039)	(0.0039)	(0.0039)	(0.0046)
$\alpha_1 = 0.40$	0.4493	0.4493	0.4493	0.4493	0.4493	0.4322
	(0.0219)	(0.0219)	(0.0219)	(0.0219)	(0.0219)	(0.0210)
$\beta_1 = 0.50$	0.4713	0.4713	0.4713	0.4713	0.4713	0.4680
	(0.0194)	(0.0194)	(0.0194)	(0.0194)	(0.0194)	(0.0200)
Log-likelihood	-2.9409e+03	-2.9409e+03	-2.9409e+03	-2.9409e+03	-2.9409e+03	-2.9735e+03
Iterations/ Generations	150	39	10	11	34	107

Table 10 Estimation results for AR(1)-ARCH(1) on the second round.

	AR(1)-ARCH(1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5627	0.5627	0.5627	0.5627	0.5627	0.5546
	(0.0059)	(0.0059)	(0.0059)	(0.0059)	(0.0059)	(0.0060)
$\phi_1 = -0.30$	-0.3285	-0.3285	-0.3285	-0.3285	-0.3285	-0.3121
	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0124)
$\alpha_0 = 0.05$	0.0505	0.0505	0.0505	0.0505	0.0505	0.0530
	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0014)
$\alpha_1 = 0.40$	0.4390	0.4390	0.4390	0.4390	0.4390	0.4165
	(0.0234)	(0.0234)	(0.0234)	(0.0234)	(0.0234)	(0.0230)
Log-likelihood	2.1800e+03	2.1800e+03	2.1800e+03	2.1800e+03	2.1800e+03	2.1775e+03
Iterations/ Generations	48	19	7	8	13	164

Table 11 Estimation results for AR(1)-GARCH(1,1) on the second round.

	AR(1)-GARCH(1,1)					
	NM	**QN**	**AS**	**SQP**	**IP**	**GA**
c= 0.55	0.5660	0.5660	0.5660	0.5660	0.5660	0.5606
	(0.0077)	(0.0077)	(0.0077)	(0.0077)	(0.0077)	(0.0077)
$\phi_1 = -0.30$	-0.3359	-0.3359	-0.3359	-0.3359	-0.3359	-0.3266
	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0122)
$\alpha_0 = 0.05$	0.0534	0.0534	0.0534	0.0534	0.0534	0.0603
	(0.0039)	(0.0015)	(0.0039)	(0.0039)	(0.0039)	(0.0044)
$\alpha_1 = 0.40$	0.4483	0.4483	0.4483	0.4483	0.4483	0.4607
	(0.0218)	(0.0248)	(0.0218)	(0.0218)	(0.0218)	(0.0217)
$\beta_1 = 0.50$	0.4726	0.4727	0.4727	0.4727	0.4727	0.4399
	(0.0193)	(0.0193)	(0.0193)	(0.0193)	(0.0193)	(0.0200)
Log-likelihood	-2.9413e+03	-2.9413e+03	-2.9413e+03	-2.9413e+03	-2.9413e+03	-2.9435e+03
Iterations/ Generations	67	22	7	8	16	85

In the case of the AR(1)-ARCH(1) and AR(1)-GARCH(1, 1) models all the algorithms, apart from the GA, agreed again, this time giving exactly the same estimates. More specifically, for the AR(1)-ARCH(1) model there was a slight improvement of the estimates, but for the AR(1)-GARCH(1, 1) model the estimates were almost the same, compared to the ones obtained before. Moreover, the GA performed well in both cases, especially in the case of the AR(1)-ARCH(1) model. Overall, in both models the estimates were good, with AS and SQP methods requiring the minimum number of iterations.

Hence, when estimating the parameters of the AR(1)-ARCH(1) or AR(1)-GARCH(1, 1) model, the choice of initial values and the choice of the algorithm is rather not important. However, when estimating the parameters of the two new nonlinear models, when using arbitrary initial values, the algorithms that perform overall best are the QN and IP, while when using initial values that are indeed close to the true parameter values, most algorithms seem to agree.

5 Applications to real data

As an illustration of the practical potential of the two new models, in this section we consider an application using financial data, in order to examine whether these nonlinear models can explain real time series data. Moreover, we use once again the AR-ARCH and AR-GARCH models as benchmarks with which to compare the results and to see if the new models can give a better fit to the series than these well-known and commonly used models. What is more, we use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for model selection, in order to help us decide which model describes the data better. According to these, the preferred model is the one with the minimum criteria values.

Our data set consists of daily figures for the Dow Jones stock price index and, more specifically, the data used are the first differences of the daily high values for the period between 3 January 2000 and 13 May 2013. The moments of the first differences of this time series can be found in Table 12 (prices in USD), while the histogram can be seen in Figure 7. The autocorrelation and partial autocorrelation plots, which indicate the use of first order

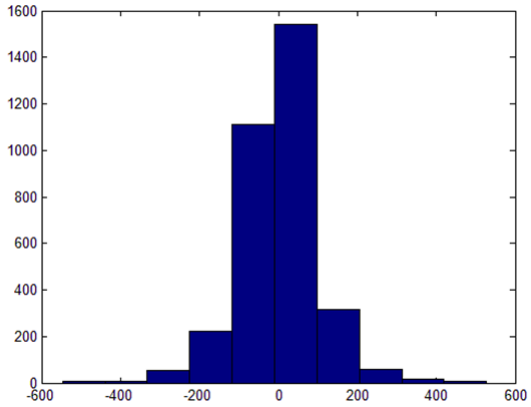


Figure 7 Histogram of the first differences of Dow Jones data.

Table 12 Moments of the first differences of Dow Jones data.

Sample size	3338
Mean	10.402
Standard deviation	966.342
Skewness	63.924
Kurtosis	-0.0545

models, can be found in Figures 8a and 8b respectively. In Table 13 there can be found the estimation results for the time series, including the maximum likelihood estimates, the standard errors, the log-likelihood value, the number of iterations required for the algorithm to converge to a solution and the AIC and BIC values for every model.

It can be seen from Table 13 that the estimates obtained for all the common parameters between the two models with ARCH errors (ExpAR-ARCH and AR-ARCH) are similar, and the estimates obtained for all the common parameters between the two models with GARCH errors (ExpAR-GARCH and AR-GARCH) are close to each other as well.

The lowest number of iterations was obtained for the AR(1)-GARCH(1, 1) model (55) and then for the ExpAR(1)-GARCH(1, 1) model (96), while the highest log-likelihood value was obtained for the ExpAR(1)-GARCH(1, 1) model (-36024) and then for the AR(1)-GARCH(1, 1) model (-36028). The lowest AIC value was given for the ExpAR(1)-GARCH(1, 1) model (72061.074) and then for the AR(1)-GARCH(1, 1) model (72065.032), while the lowest BIC value was obtained for the AR(1)-GARCH(1, 1) model (72095.69) and then for the ExpAR(1)-GARCH(1, 1) model (72103.866). Hence, according to the Akaike Information Criterion, the preferred model is the ExpAR(1)-GARCH(1, 1), while according to the Bayesian Information Criterion the preferred model is the AR(1)-GARCH(1, 1). However, we should bear in mind that the latter result could be due to the fact that the BIC penalises a higher number of parameters more than the AIC.

Table 14 shows the t-statistics for the estimates of the ExpAR(1)-GARCH(1,1) model. It is worth noting that the estimates of all the parameters for the preferred model, according to the AIC, are statistically significant at a 5% level.

Table 15 shows the values of the LR tests, according to which under the null hypothesis the true model is either the AR(1)-ARCH(1), or the ExpAR(1)-ARCH(1) or the AR(1)-GARCH(1,1) against the alternative that the true model is the ExpAR(1)-GARCH(1,1). According to the results, we can reject the null hypothesis that the true model is the AR(1)-

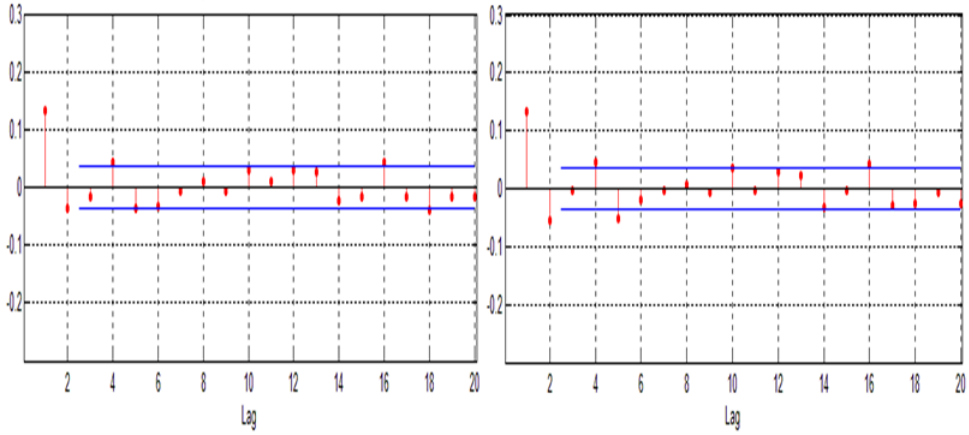


Figure 8 Autocorrelation (a) and Partial Autocorrelation (b) plot of the first differences of Dow Jones data.

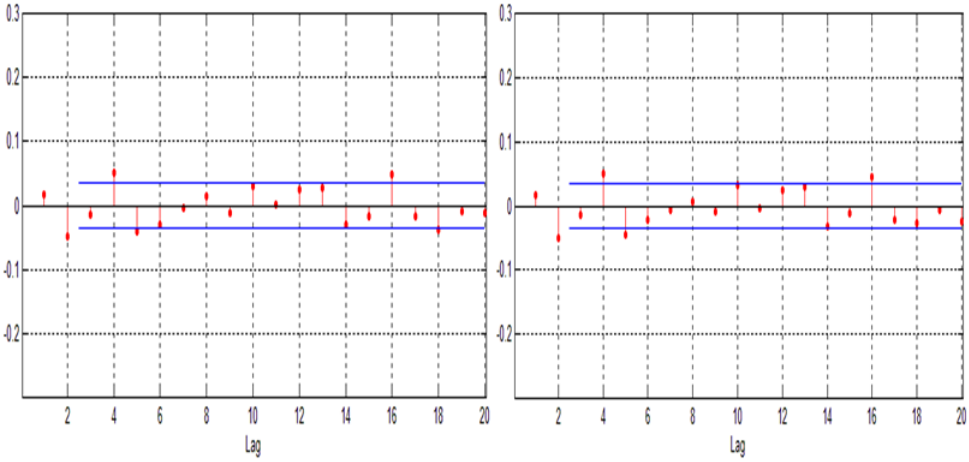


Figure 9 Autocorrelation (a) and Partial Autocorrelation (b) plot of the residuals of the ExpAR(1)-GARCH(1,1) model.

ARCH(1) or the ExpAR(1)-ARCH(1) at a 1% level. We can also reject the null hypothesis that the true model is the AR(1)-GARCH(1,1) model at a 5% level. Consequently, at a 5% level we can accept the alternative hypothesis that the true model is the ExpAR(1)-GARCH(1,1), which verifies the model selection according to the AIC.

In addition, Figure 9 shows the autocorrelation and partial autocorrelation plots of the residuals of the estimated ExpAR(1)-GARCH(1,1) model, which are useful tools to assess the presence of autocorrelation at individual lags. According to these plots, most sample autocorrelations and partial autocorrelations fall inside the 95% confidence bounds and change sign indicating the residuals to be random. Hence, the choice of the ExpAR(1)-GARCH(1,1) model for this time series seems to be appropriate.

6 Conclusions

In this paper two new nonlinear time series models, namely the ExpAR-ARCH and ExpAR-GARCH, have been suggested. Simulated series have been shown and several methods have

Table 13 Estimation results for the first differences of Dow Jones data.

	ExpAR(1)-ARCH(1)	**ExpAR(1)-GARCH(1,1)**	**AR(1)-ARCH(1)**	**AR(1)-GARCH(1,1)**
c	4.0217	4.8205	4.0059	4.6790
	(1.1359)	(0.9304)	(1.1487)	(0.9163)
ϕ_1	0.1072	0.1220	0.1072	0.1233
	(0.0142)	(0.0127)	(0.0142)	(0.0127)
π_1	210.2257	3.0392	-	-
	(37.4497)	(1.3967)	(-)	(-)
γ	9.3089	0.0231	-	-
	(4.0451)	(0.0095)	(-)	(-)
α_0	7.6998e+03	69.1401	7.7063e+03	69.5382
	(6.1816)	(9.0288)	(7.3042)	(5.4370)
α_1	0.1638	0.0580	0.1632	0.0580
	(0.0162)	(0.0046)	(0.0162)	(0.0045)
β_1	-	0.9344	-	0.9344
	(-)	(0.0046)	(-)	(0.0041)
Log-likelihood	-3.6747e+04	-3.6024e+04	-3.6748e+04	-3.6028e+04
Iterations	241	96	133	55
AIC	73505.116	72061.074	73503.804	72065.032
BIC	73541.795	72103.866	73528.33	72095.69

Table 14 t-statistics for the estimates of the ExpAR(1)-GARCH(1,1) model.

	c	ϕ_1	π_1	γ	α_0	α_1	β_1
Estimates	4.8205	0.1220	3.0392	0.0231	69.1401	0.0580	0.9344
	(0.9304)	(0.0127)	(1.3967)	(0.0095)	(9.0288)	(0.0046)	(0.0046)
t-statistic	5.1811	9.6063	2.1760	2.4316	7.6577	12.6087	203.1304

Table 15 Likelihood Ratio tests, (*): reject null hypothesis at $\alpha = 0.01$, (**): reject null hypothesis at $\alpha = 0.05$, but not at $\alpha = 0.01$.

	AR(1)-ARCH(1)	**ExpAR(1)-ARCH(1)**	**AR(1)-GARCH(1,1)**
ExpAR(1)-GARCH(1,1)	1448 (*)	1446 (*)	8 (**)

been used and compared in the estimation, showing that the algorithms that performed better when using arbitrary initial values are the Quasi-Newton and the Interior Point, while most algorithms gave similar results when using initial guesses that are indeed close to the true parameter values. The results have been compared to the AR-ARCH and AR-GARCH models, in the case of which the choice of initial values or of algorithm is surprisingly not so important.

It has also been shown that the new models, and in fact low orders of them, can describe specific financial time series data. In addition, according to the Akaike Information Criterion, the ExpAR-GARCH model can even fit better than the well-known and widely used AR-ARCH and AR-GARCH models. Furthermore, some diagnostic tests have been used in order to verify our model selection.

All in all, the ExpAR-ARCH and ExpAR-GARCH models can be a useful tools in describing nonlinear behaviour in financial time series and have the potential of describing and fitting various real time series data. It should be noted that our suggested models can be extended by allowing for other forms of conditional variance. What is more, it is of interest to apply the ExpAR model with conditional heteroscedastic errors to other important financial and economic time series and to check the new models' forecasting performance as well. This will be the purpose of future investigation.

Acknowledgements. The author would like to thank her supervisor, Prof. Terence Mills, for his insightful comments.

References

1 A. Amendola and M. Niglio. *Non-linear Dynamics and Evaluation of forecasts using High-Frequency Time Series*. Quaderni di Statistica 2, 157–171, 2000.

2 T. Bollerslev *Generalized autoregressive conditional heteroskedasticity*. Journal of econometrics 31(3), 307–327, 1986.

3 G. E. Box and G. M. Jenkins. *Times series analysis: Forecasting and Control*. San Francisco: Holden Day, 1970.

4 G. Busetti and M. Manera. *STAR-GARCH Models for Stock Market Interactions in the Pacific Basin Region, Japan and US*. Nota Di Lavoro 43, 2003.

5 K. S. Chan and H. Tong. *On estimating thresholds in autoregressive models*.Journal of Time Series Analysis 7, 179–90, 1986.

6 T. C. Chiang and S.-C. Doong. *Empirical Analysis of Stock Returns and Volatility: Evidence from Seven Asian Stock Markets Based on TAR-GARCH Model*. Review of Quantitative Finance and Accounting 17, 301–318, 2001.

7 R. F. Engle. *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*. Econometrica: Journal of the Econometric Society, 987–1007, 1982.

8 C. W. J. Granger and A. P. Andersen. *An introduction to bilinear time series models*. Vandenhoeck und Ruprecht, Gottingen,1978.

9 G. Koutmos. *Do emerging and developed stock markets behave alike? Evidence from six pacific basin stock markets*. Journal of International Financial Markets, Institutions and Money 7, 221–234, 1997.

10 B. LeBaron. *Some relations between volatility and serial correlations in stock market returns*. The Journal of Business 65(2), 199–219, 1992.

11 Y.N. Lee and W.K. Li. *On smooth transition double threshold models*. Statistics and Finance: An Interface, 205–225, 1998.

12 C. W. Li and W. K. L.i *On a double threshold autoregressive heteroscedastic time series model*. Journal of Applied Econometrics 11, 253–274, 1996.

13 W. K. Li and K. Lam. *Modelling Asymmetry in Stock Returns by a Threshold Autoregressive Heteroscedastic Model*. Journal of the Royal Statistical Society, Series D (The Statistician) 44(3), 333–341, 1995.

14 S. Lundbergh and T. Terasvirta. *Modelling economic high-frequency time series with STAR-STGARCH models*. Tinbergen Institute, 1999.

15 M. P. Munoz, M. D. Marquez and L. M. Acosta. *Forecasting Volatility by Means of Threshold Models*. Journal of Forecasting 26, 343–363, 2007.

16 M. Osinska and M. Witkowski. *The TAR-GARCH Models with Application to Financial Time Series*. Dynamic Econometric Models 6, 105–116, 2004.

17 T. Ozaki. *Non-linear time series models for non-linear random vibrations*. Journal of Applied Probability 17, 84–93, 1980.

18 J. H. Stock. *Estimating continuous-time processes subject to time deformation: An application to postwar U.S. GNP*. Journal of the American Statistical Society 83, 77–85, 1988.

19 H. Tong. *Some comments on the Canadian lynx data (with discussion)*. Journal of the Royal Statistical Society. Series A (General) 140, 432–436, 1977.

20 H. Tong. *Non-linear time series: a dynamical system approach*. Oxford University Press, 1990.

21 H. Tong and K. S. Lim. *Threshold autoregression, limit cycles and cyclical data*. Journal of the royal Statistical Society, Series B (Methodological) 42(3), 245–292, 1980.

22 G. U. Yule. *On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers*. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 226, 267–298, 1927.

Coordinating push and pull flows in a lost sales stochastic supply chain

Georgios Varlas and Michael Vidalis

Department of Business Administration, University of the Aegean
Michalon 8, Chios 82100, Greece
g.varlas@aegean.gr, m.vidalis@aegean.gr

Abstract

In this paper a serial, three echelon, push-pull supply chain is investigated. The supply chain consists of a provider, a distribution centre (buffer) and a retailer. The material flow between upstream stages is push type, while between downstream stages it is driven by continuous review, reorder point/order quantity inventory control policy. Exponentially distributed lead times between stages are assumed. External demand occurs according to pure Poisson, while the demand that cannot be met is lost. The system is modelled using matrix analytic methods as a Markov birth-and-death process. An algorithm is developed to generate the transition matrix for different parameters of the system. Then, the corresponding system of stationary linear equations is generated and the solution of the stationary probabilities is provided. Key performance metrics such as average inventories and customer service levels at each echelon of the system can be computed. The algorithm is programmed in Matlab© and its validity is tested using simulation, with the two approaches giving practically identical results. The contribution of our work is an exact algorithm for a lost sales push-pull supply network. This algorithm can be used to evaluate different scenarios for supply chain design, to explore the dynamics of a push-pull system, or as an optimization tool.

1998 ACM Subject Classification G.3 Probability and Statistics

Keywords and phrases Supply Chain Management, Push-Pull systems, Markov Processes

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.52

1 Introduction and Literature Review

The supply chain (SC) consists of all the parties involved in manufacturing, distribution, and delivery of the product to customers. Members of various echelons in the SC are related to each other, either directly or through intermediaries. Decisions made in any echelon of the SC can affect the costs of other members and vice versa. Supply chain management (SCM) involves the management of flows between the stages of a supply chain, so as to maximize total expected profitability. Six tool drivers may be used to improve supply chain performance: Inventory, Transportation, Facilities, Information, Sourcing, and Pricing [4].

Inventory control plays an important role in supply chain management. Properly controlled inventory can satisfy customers' demand, smooth the production plans, and reduce the operational costs. In practice, inventory control systems usually operate in dynamic environments. Calculating the exact ordering quantity, deciding the proper reordering point, choosing the right inventory reviewing policy, and managing the safety stock are key factors for the SC profitability.

The reordering process is characterized by the review interval, the determination of the order size, the order costs and the objective function. Two types of review systems are widely used in business and industry. Either inventory is continuously monitored (continuous

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 52–62

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

review), or inventory is reviewed at regular periodic intervals of length R (periodic review). Whether or not to order at a review instant is usually determined by a reorder level denoted by s. This is the inventory position at which a vendor is triggered to place a replenishment order so as to maintain an adequate supply of items to accommodate current and new customers.

Common assumptions of the inventory model to represent the system concern the demand distribution, the lead time (deterministic or stochastic), and the maximum number of outstanding orders. Another important characteristic concerns the demand that cannot be met by the inventory on hand. Such demand can be back-ordered to be met in the future, or can be lost (lost sales assumption). Latter cases appear to be more difficult to analyse and such models have received less attention in the literature [2].

Usually, a Markov model, based on the characteristics of the system and the assumptions made, is developed to represent the on-hand inventory level and the individual outstanding orders. The decision (or ordering) points in such a model are the time instants at which either a demand occurs and no order is outstanding, or a replenishment order is delivered. Based on the transition probabilities and steady-state behaviour of the system, the long-run behaviour of the inventory model is analysed. The stationary distribution function of the on-hand inventory level is used to analyse the inventory system in terms of expected average cost and service level. The pioneer work in this area was that of Clark and Scarf [5] who considered an inventory system with periodic review using echelon stock policy.

In the final phase, the inventory control variables, such as the reorder level and order quantities, are set. Either an exact procedure or an approximation procedure can be used to find these values. Two types of exact procedures are commonly used in literature, namely a policy iteration algorithm and an extensive numerical search procedure.

In general, production/inventory systems can be classified as push, pull, or hybrid push/pull-type systems. In a push-type system the parts are released to the next station as quickly as possible to avoid starvation of the downstream stations. On the other hand, the pull-type system drives production based upon customer demand. Such systems are widely used and different modelling approaches have been proposed. Chen [3] generalizes the Clark and Scarf model by allowing batch transfers of inventories in a serial network with n stages and backorders. Badinelli [1] constructs a model of the steady-state values of on-hand inventory and backorders for each facility of a serial inventory system, where each facility follows a (Q, R) policy based on installation stock. The descriptive model he presents is intended for optimizing the parameters of such a policy and for obtaining theoretical results about the behaviour of the system. Finally, Gupta and Selvaraju [9] study the effect of stock allocation among different stages, when the total amount of stock in the supply system is fixed at the optimal level. They develop an approximation scheme for performance evaluation of serial supply systems when each stage manages its planned inventories according to a base-stock policy.

In the hybrid push/pull system the production at the earlier upstream stations is push-type, while the production of the later downstream stations is controlled by pull-type policies. The push-pull boundary or junction point is defined as the last push station and determines which stations are push systems and which stations are pull systems. In most cases hybrid systems perform better than pure push, or pure pull systems, while they are more flexible to address growing product variety and shorter product life cycles. However, their analysis is more complicated. Cochran and Kim [6] study with Simulated Annealing a horizontally integrated hybrid production system (HIHPS) with a movable junction point. Their proposed solutions include the location of the junction point, the safety stock level, and the number of

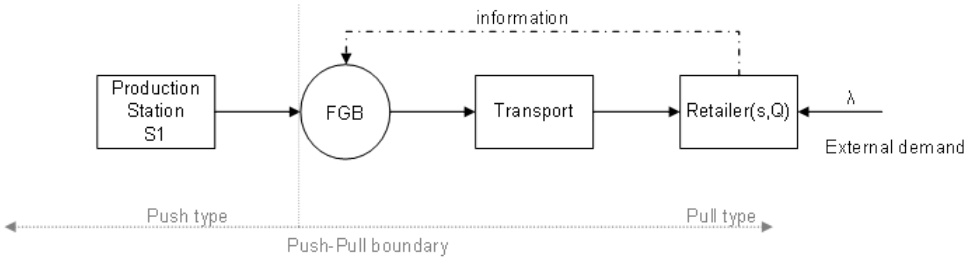


Figure 1 System layout.

kanbans needed in the pull system. Ghrayeb et. al [8] investigate a hybrid push/pull system of an assemble-to order manufacturing environment. They use discreet event simulation along with a genetic algorithm and the objective function for their model is to minimize the sum of inventory holding cost and delivery lead time cost. Finally, Cuypere et. al [7] introduce a Markovian model for push-pull systems with backlogged orders, basing their analysis on quasi birth-and-death processes.

The main goal of our work is to provide an algorithm for the exact evaluation of a push-pull supply network with lost sales. The resulting descriptive model can be used as a design tool or as a tool for the optimization of the parameters of the system.

2 Description of the System

In this article a single product, linear, push-pull supply chain is investigated. The system under consideration is shown in Figure 1. A reliable station S_1 produces (or administers in the system) product units at a rate μ_1 and exponentially distributed inter-arrival times. Finished products are stored in a finite Finished Goods Buffer (FGB). Inventory at buffer is denoted by B_t. In the case where S_1 completes processing, but on completion FGB is full, station S_1 blocks (blocking after processing). Station S_1 consists the push section of the system. Downstream, the retailer R holds inventory I_t and faces external demand with pure Poisson characteristics (customers' inter-arrival times are exponentially distributed and every customer asks for exactly one unit). When the retailer is out of stock, occurring demand is lost. The retailer follows continuous review inventory control policy with parameters (s,Q). When inventory I_t reaches the reorder point s, a replenishment order of Q units is placed on the buffer. The actual level of the sent order depends on the available inventory at buffer. If $B_t \geq Q$, a full order is dispatched to the retailer. Otherwise, an incomplete order is dispatched. In the case where FGB is empty, dispatching is suspended until one unit finishes processing at S_1, upon which it is immediately forwarded for transportation to the retailer. Transportation is modelled as a virtual station T. Inventory in transit is denoted by T_t. In the model, transportation is considered independent from both FGB and retailer. On transportation initiation inventory T_t is subtracted from the buffer and remains in the virtual station T until on transportation completion it is added to the inventory of the retailer I_t. Exponentially distributed times for the transportation are assumed.

To model the system, the following assumptions are made:

1. Both customer demand and lead time are stochastic.
2. There are no back-orders. Demand that cannot be met from inventory on hand is lost both at the retailer and the buffer.
3. At any given time only one order can be in transit from FGB to the retailer.

4. The retailer follows continuous review inventory control policy with parameters (s,Q). Decision variables of the retailer are reorder point s and order quantity Q. In other words, the retailer's problem is the optimization of s and Q simultaneously.
5. Order quantity Q is constant.
6. Transportation is modelled as an independent station and inventory in transit depends on B_t at the time of transportation initiation.
7. Station S_1 never starves.
8. Station S_1 blocks when on completing the processing of a unit, finds the FGB full (blocking after processing). In the case where Q is greater than buffer capacity, the blocked unit is considered available for transportation to the retailer along with the inventory of the buffer (The blocked unit is considered part of the buffer). Otherwise, the blocked unit is transferred to the buffer immediately after there is available space and at the same time station S_1 resumes production.
9. There are no loading/unloading times.
10. All stations are reliable.
11. All times are exponentially distributed.
12. For methodology reasons, it is assumed that no two events can occur at exactly the same time.

3 Description of the Model

3.1 States definition

The system cannot be modelled as a quasi birth-and-death (QBD) process since the assumption of (s,Q) policy allows for transitions between non-adjacent levels. However, the system can be modelled as a continuous time-discreet space Markov process using matrix analytic methods. Taking advantage of repeating structures, an algorithm is developed to generate the transition matrix for different parameters of the system. Then, the corresponding system of stationary linear equations is generated and the solution of the stationary probabilities is provided. Using stationary probabilities, key performance metrics, such as average inventories and service levels at each echelon of the system can be computed.

The design variables that determine the dimension and structure of the transition matrix are:
B: The capacity of the finished goods buffer (FGB).
s: The reorder point at the retailer.
Q: The quantity of the orders requested by the retailer.
All three variables are assumed to be positive integers or zero, with the exception of Q which obviously cannot be zero. Although some scenarios lack physical meaning (for example when $Q > B + 1$), for the development of the algorithm no assumptions about the variable values are made. The other parameters of the model are:
μ_1: The production rate of the Station 1, or rate of units admission in the system (exponential inter-arrival times)
μ_2: The transfer rate of a replenishment order from the buffer to the retailer (exponential times)
λ: The rate of external customers' arrivals (pure Poisson demand)

We will illustrate the methodology using a simple example. We assume buffer capacity $B = 2$, reorder point $s = 1$ and order quantity $Q = 2$. At any moment t, the state of the system can be defined by a three dimensional vector (B_t, T_t, I_t), where:
B_t: The level of inventory on hand at the FGB. $0 \leq B_t \leq B + 1$, where the case $B_t = B + 1$ corresponds to blocking (see assumption 8). In our example $0 \leq B_t \leq 3$.

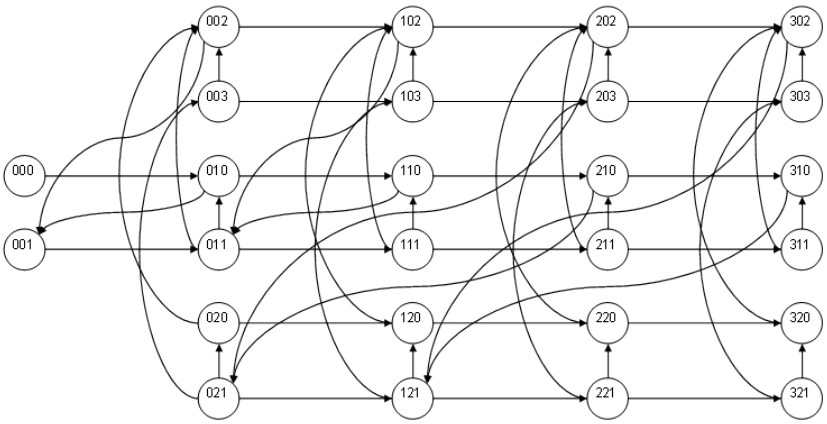


Figure 2 State transition diagram for $B = 2$, $s = 1$, $Q = 2$.

T_t : The number of product units in transit from FGB to the retailer. $0 \leq T_t \leq Q$. $T_t = 0$ means that there is no inventory in transit, while when $T_t = Q$ we have a complete order in transit to the retailer. $0 < T_t < Q$ corresponds to incomplete order. In our example $0 \leq T_t \leq 2$

I_t : The inventory on hand at the retailer. In general $0 \leq I_t \leq s + Q$. In our example $0 \leq I_t \leq 3$

The state space S of the Markov process is comprised of all the possible triplets (B_t, T_t, I_t) and its dimension depends on B, s and Q. For the example under consideration there are 26 possible states. It can be easily proved that for any value of the given parameters, the dimension of the state space is given by

$$N_B^{s,Q} = (s+1) + (s+2) \cdot Q \cdot (B+2)$$

3.2 State transitions

The state of the system can be altered instantaneously by three kinds of events.

1. The completion of processing of one product unit at station S_1. In this case B_t increases one unit. In infinitesimal time dt, the possibility of the event occurring is $\mu_1 \cdot dt + o(dt)$. $o(dt)$ is an unspecified function such that $\lim_{dt \to 0} \frac{o(dt)}{dt} = 0$.
2. The arrival of an outstanding order at the retailer. In this case the inventory on hand of the retailer I_t increases by T_t units. If the new value of I_t is above the reorder point, then T_t resets to zero. Otherwise, a new transfer from FGB is initiated. T_t takes the value of the new order and B_t decreases correspondingly. In infinitesimal time dt, the possibility of the event occurring is $\mu_2 \cdot dt + o(dt)$.
3. The occurrence of external demand. In this case the inventory on hand of the retailer decreases by one unit. If the new inventory equals the reorder point s, a replenishment order is given to the FGB. T_t takes the value of inventory in transit and B_t decreases correspondingly. In infinitesimal time dt, the possibility of the event occurring is $\lambda \cdot dt + o(dt)$.

In Figure 2 is given the state transition diagram for the example under consideration. There are certain symmetries in the diagram. Such symmetries are also present in the transition matrix and form the basis for the development of the algorithm which is the target of our analysis.

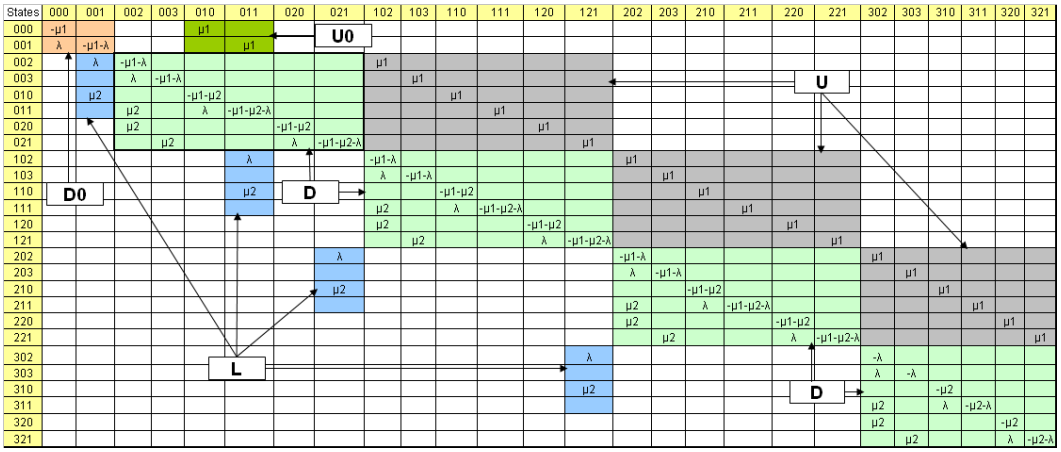


Figure 3 Transition Matrix for $B = 2$, $s = 1$, $Q = 2$.

3.3 The Transition Matrix

Before displaying the transition matrix, we must define a linear ordering of the states. We use the lexicographical ordering [10]. We take as basic level the subset of all states corresponding to a fixed buffer inventory B_t. Such levels correspond to columns in the state transition diagram. Within each level the states are grouped according to the inventory in transit T_t. For fixed level and fixed inventory in transit, the states are ordered by inventory at retailer I_t. To summarize: State (x, y, z) precedes state (x', y', z') if x<x'; State (x, y, z) precedes state (x, y', z') if y<y'; State (x, y, z) precedes state (x, y, z') if z<z'. The transition matrix for the example under consideration is given in Figure 3.

The transition matrix can be divided into sub-matrices with well defined and predictable characteristics. Some examples of constituent sub-matrices are given in Figure 4.

The first diagonal sub-matrix D_0 corresponds to the boundary states where there is no inventory in transit, $I_t < s$ and $B_t = 0$. In our example it is a 2x2 block at the top left. In general it is a $(s+1) \times (s+1)$ sub-matrix. On the diagonal, the basic repeating block D is a $(s+2) \cdot Q \times (s+2) \cdot Q$ sub-matrix. It corresponds to analogous transitions for different values of buffer inventory (levels), and it is repeated B+2 times. D can be further analysed into constituent sub-sub-matrices.

- D_1 is a $Q \times Q$ block on the diagonal of D. It corresponds to the diagonal of transition matrix P, where no event occurs, and to the occurrence of external demand when no replenishment order is initiated ($I_t > s$)
- D_2 is a $(s+1) \times (s+1)$ block also on the diagonal of D. Within each D block, D_2 is repeated Q times. It corresponds to the diagonal of transition matrix P and to the occurrence of external demand when no replenishment order is initiated ($T_t > 0$).
- D_3 is a $k \cdot (s+1) \times Q$ block, where k=min(s,Q). It is located just below D_1 and corresponds to the arrival of replenishment orders at the retailer when the new I_t exceeds s. D_3 consists of k blocks of s+1 lines. ith block consists of s+1-i zero lines (corresponding to the arrival of a replenishment order when the new $I_t \leq s$) and a left aligned $i \times i$ diagonal matrix of μ_2.
- D_4 occupies the left down corner of D. It occurs only when $Q > s$ and corresponds to replenishment orders where $T_t > s$. D_4 is a $(s+1) \cdot f \times Q$ sub-matrix, where f= max(Q-s,0). It can be divided to Q-s, left aligned diagonal blocks of μ_2. Each block has dimension $(s+1) \times (s+1)$ and each subsequent block is located one column to the right.

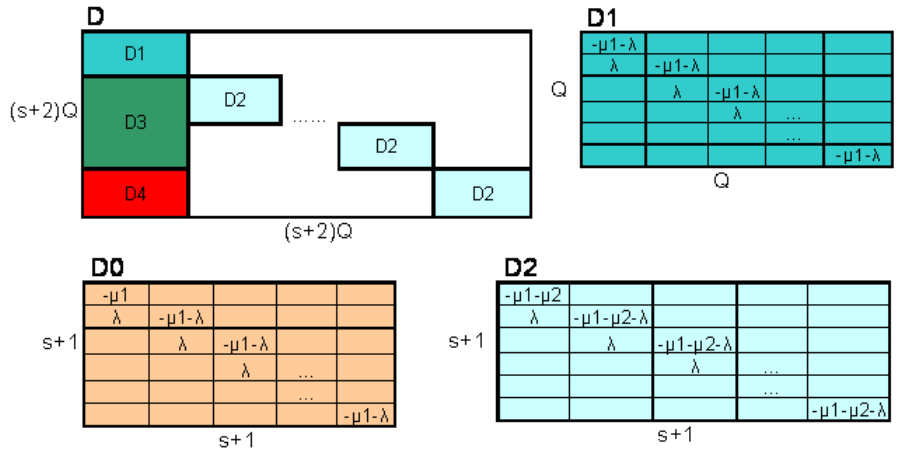


Figure 4 Examples of sub-matrices.

The last sub-matrix D corresponds to the boundary states where $B_t = B + 1$. In these states station S_1 is blocked and therefore there can be no arrivals at buffer. Consequently $-\mu_1$ is subtracted from the diagonal elements of the last D sub-matrix.

Upper diagonal sub-matrices correspond to arrivals from S_1 to buffer. Since S_1 processes one unit at a time, only transitions to adjacent levels occur. The first upper diagonal sub-matrix U_0 corresponds to arrivals at buffer, while the system is at the boundary states where $T_t = 0$, $I_t < s$ and $B_t = 0$. In such cases, the arriving unit is immediately forwarded for transportation to the retailer. U_0 is a $(s + 1) \times (s + 1)$ diagonal block of μ_1. U is the repeating upper diagonal block. It is a $(s + 2) \cdot Q \times (s + 2) \cdot Q$ diagonal matrix of μ_1. It is repeated B+1 times, corresponding to different levels.

The repeating block L below the diagonal describes transitions where there is triggering of replenishment orders. The events that can trigger replenishment orders are a) the occurrence of external demand while $I_t = s + 1$ and b) the arrival of a replenishment order when the updated I_t does not exceed s. For s=0, $L = \lambda$. For s>0, L is a $(s + 1) \cdot k + Q \times s$ block, where k=min(s,Q). For such cases, block L has the following structure: The element (1,s)=λ; then follow Q-1 zero lines; then follow k, $(s + 1) \times s$ blocks where the ith block is a $(s + 1 - i) \times (s + 1 - i)$ right aligned diagonal matrix of μ_2 followed by i zero lines.

L is the most complicated sub-matrix since it corresponds to transitions between non-adjacent levels. Sub-matrix L occurs B+2 times and its exact positioning in the transition matrix P depends on the parameters B,s and Q. The first occurrence of L corresponds to the boundary conditions where the buffer is empty. The next h = min(B+1,Q) occurrences correspond to cases where $B_t \leq Q$. Finally, the next B+1-h occurrences correspond to initiation of complete orders when $B_t > Q$.

3.4 Performance Measures

From the transition matrix, the corresponding system of linear equations can be determined and the vector of stationary probabilities can be computed. Using the stationary probabilities, we can calculate performance measures for the system under consideration. Again, we take advantage of the structure of the transition matrix. To illustrate through an example, for $B = 2$, $s = 1$, $Q = 2$, the percentage of external demand that is met from inventory on hand:

$$\begin{aligned} \text{Fill Rate retailer} &= P(I_t > 0) = 1 - P(I_t = 0) = \\ &= 1 - (\pi_{000} + \pi_{010} + \pi_{020} + \pi_{110} + \pi_{120} + \pi_{210} + \pi_{220} + \pi_{310} + \pi_{320}) . \end{aligned}$$

In general, taking advantage of the transition matrix structure,

$$FillRate = 1 - \pi_1 - \sum_{j=0}^{B+1} \sum_{i=0}^{Q-1} \pi_{r+i\cdot(s+1)} ,$$

where $r = s + Q + 2 + j \cdot (s + 2) \cdot Q$, and π_i is the ith element of the stationary probability vector π. It is reminded that the sequence of system states is defined according to certain rules as expounded earlier.

Similarly, for the average inventory at buffer (Work in Process Buffer or WIP Buffer), including blocked units:
Inventory buffer $= c \otimes b$, where:
c: is a line vector with the possible positive values of B_t In our example c= [1, 2, 3], and generally c=[1, 2, . . . ,B+1].
b: is a column vector with the ith element giving the probability that i units belong to the buffer. In our example $b^\tau = [(\pi_{102} + \pi_{103} + \pi_{110} + \pi_{111} + \pi_{120} + \pi_{121})(\pi_{202} + \pi_{203} + \pi_{210} + \pi_{211} + \pi_{220} + \pi_{221})(\pi_{302} + \pi_{303} + \pi_{310} + \pi_{311} + \pi_{320} + \pi_{321})]$

From the matrix structure, and especially from the levels we have defined according to B_t, it can be inferred that for the j+1 element of vector b

$$b_{j+1} = \sum_{i=r}^{r+(s+2)\cdot Q-1} \pi_i ,$$

where $r = (s + 2)(1 + Q + j \cdot Q)$
so that b can be calculated and average WIP buffer can be computed as the product $c \otimes b$.

In a similar way we can also calculate the rest performance measures of concern, namely average inventory at retailer (WIP retailer), percentage blocked for S_1, and average inventory in transit from the buffer to the retailer (WIP in transit).

4 Verification

The validity of the algorithm is verified using simulation. A simulation model of the system described in section 2 is developed using Arena© simulation package. An approach similar to the cycle view of supply chains is adopted. The system is modelled using three cycles corresponding to the interfaces between Buffer, Transportation and Retailer. The results of the algorithm described in section 3 are collated with simulation results for the same system parameter values and the two approaches are found to give practically identical results. A simulation time of 1000000 time units was selected as it was deemed long enough to provide statistically vigorous results. Moreover, a warm-up period of 20000 time units was selected, so as to eliminate the effect of the initial conditions. Figure 5 gives the comparison of analytic and simulation solutions for performance measure Fill Rate across various scenarios. The parameters of the system are: $\mu_1 = 1$, $\mu_2 = 0.5$, $\lambda = 1$, $0 \leq B \leq 10$, $0 \leq s \leq 10$, and $0 \leq Q \leq 11$. % Fill Rate Difference $= \frac{FR_{Simulation} - FR_{Analytic}}{FR_{Analytic}}$. The difference does not exceed 0.5 %, well within the limits of the expected variability due to the statistical nature of simulation results.

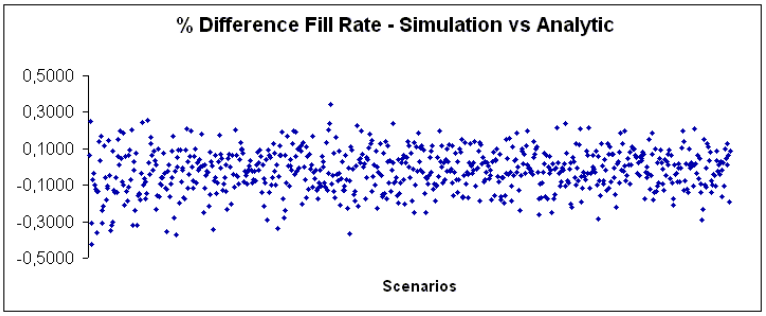


Figure 5 Simulation vs Analytic results for various scenarios.

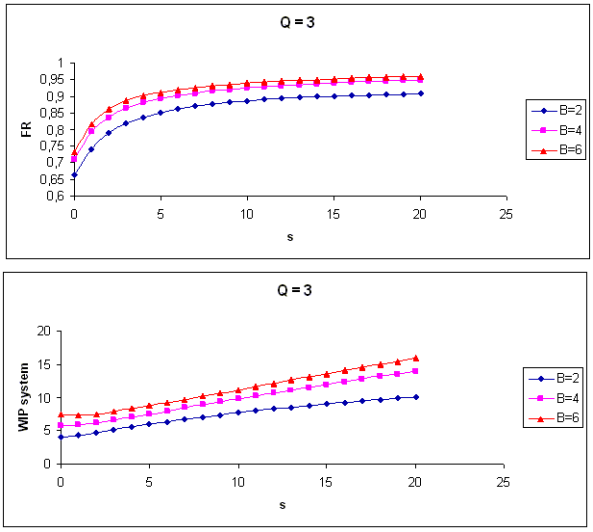


Figure 6 The effect of s on Fill rate and WIP system for different B levels.

5 Results

We use the analytic model to investigate the effect of each parameter B, s and Q on the performance measures of the system. We choose a balanced system, where $\mu_1 = \mu_2 = \lambda = 1$. Costs in a supply chain are mainly associated with holding inventories and lost sales, so we focus our analysis on the performance measures Fill Rate and average inventory in the system (WIP System). In push-type systems costs are also associated with blocking, in terms of lowered utilization and production disruptions, but at this stage blocking is not investigated. Figure 6 gives the effect of s on the performance measures for different levels of B and constant Q. For a given level of B, increasing s causes an almost linear increase in the average system inventory (Work in Process, or WIP system). At the same time, the improving effect on fill rate diminishes with increasing s. From the model it is inferred that beyond a point, changing the value of s would not be an advisable strategy for the improvement of supply chain performance.

Figure 7 gives the effect of B on the performance measures of concern for different levels of Q and constant s. We can see that for a range of values of B there are different (B,s,Q) policies yielding different levels of Fill Rate for approximately the same level of average system inventory (WIP system). In such cases there is a potential of enhancing supply chain

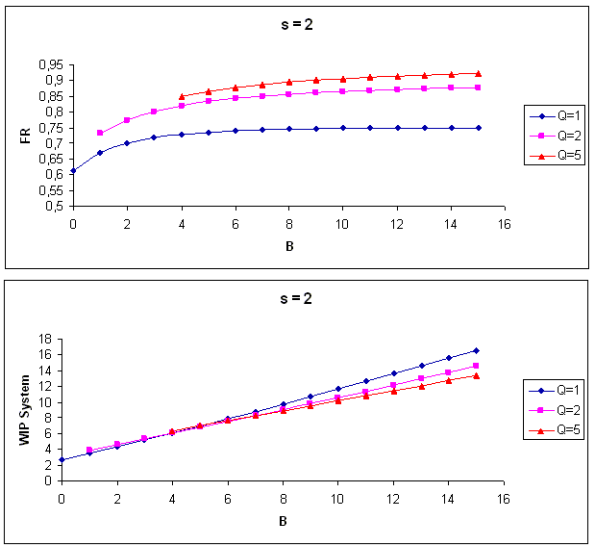


Figure 7 The effect of B on Fill Rate and WIP system for different levels of Q.

performance without incurring further costs. In a real situation, decision makers should further experiment in order to define the optimal policy within the given constraints. It should be noted that a change of policy would have different effects on the different members of the supply chain. For example, for $B = 5$, changing the value of Q from 1 to 5 would increase Fill rate significantly, but the average inventory (and thus the corresponding costs) at the retailer would also increase. However, such an increase would be compensated by the decrease of the average inventory at the buffer so that the average inventory in the system remains the same. On the whole, the performance of the supply chain can be improved, but a "global" viewpoint and centralised decision making would be required.

6 Conclusions – Future research

In our work we developed an exact algorithm for the analysis of a simple, serial, push-pull supply chain. The proposed descriptive model captures relationships between variables, offers insight on key features of the system at hand, and can be used as a design tool for the evaluation of appropriate systems and the determination of their optimal characteristics. By extensive enumeration and evaluation of the possible policies (B,s,Q), the optimal policy that will minimize average system inventory for a given threshold value of Fill Rate, or that will maximize Fill Rate for a given maximum average system inventory, can be determined.

As indicated by the results for a balanced system, all parameters B, s and Q can have an impact on system performance. The relative importance of each depends on the specific range of its values. With regard to the average inventory in the system, Q seems to have a lesser effect since changes in the average inventory at the retailer are counterbalanced by the changes in the average buffer inventory. On the other hand, an increase in s or B causes an almost linear increase in average system inventory. With regard to Fill Rate, increasing any of the parameters B,s,and Q improves system performance. However, changing only one of the parameters, there is a threshold Fill Rate value that cannot be exceeded, and the effect of each parameter diminishes as this value is approached. Due to the dynamic nature of the system, effective decision making should take into account the effect of each policy on the whole supply chain, since a local view may lead to sub-optimal solutions.

In a further step of our research, the algorithm can be expanded to include different demand characteristics (for example compound Poisson) or longer chains. Members may be added either upstream (push segment), or downstream (pull segment). The use of phase type distributions (Erlang, Coxian) instead of exponential distribution could also be a possible object of further research.

References

1 R. D. Badinelli. A model for continuous-review pull policies in serial inventory systems. *Operations Research*, 40(1):142–156, 1992.

2 M. Bijvank and I. Vis. Lost-sales inventory theory: A review. *European Journal of Operational Research*, 215:1–13, 2011.

3 F. Chen. Optimal policies for multi-echelon inventory problems with batch ordering. *Operations Research*, 48(3):376–389, 2000.

4 S. Chopra and P. Meindl. *Supply Chain Management, Strategy, Planning & Operations, 3rd edition*, 44–72. Pearson International, 2007.

5 A. J. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490, 1960.

6 J. K. Cochran and S. S. Kim. Optimum junction point location and inventory levels in serial hybrid push/pull production systems. *International Journal of Production Research*, 36(4):1141–1155, 1998.

7 E. Cuypere, K. Turck and D. Fiems. A Queueing Theoretic Approach to Decoupling Inventory. Analytical and Stochastic Modeling Techniques and Applications. In proceedings *19th International Conference, ASMTA 2012, Grenoble, France*. 150–164, 2012.

8 O. Ghrayeb, N. Phojanamongkolkij and B. A. Tan. A hybrid push/pull system in assemble-to-order manufacturing environment. *Journal of Intelligent Manufacturing*, 20(4):379–387, 2009.

9 D. Gupta and N. Selvaraju. Performance Evaluation and Stock Allocation in Capacitated Serial Supply Systems. *Manufacturing and Service Operations Management*, 8(2):169–191, 2006.

10 G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on statistics and Applied probability, 1999.

Mathematical Programming bounds for Large-Scale Unit Commitment Problems in Medium-Term Energy System Simulations

Alberto Ceselli[1], Alberto Gelmini[2], Giovanni Righini[1], and Andrea Taverna[2,3]

1 **Università degli Studi di Milano**
Dipartimento di Informatica, Via Bramante 65 – 26013 Crema (CR), Italy
{alberto.ceselli,giovanni.righini}@unimi.it
2 **Ricerca sul Sistema Energetico – RSE S.p.A., Via Rubattino 54 – 20134 Milano, Italy**
alberto.gelmini@rse-web.it
3 **Università degli Studi di Milano**
Dipartimento di Matematica, Via Saldini 50 – 20133 Milano, Italy
andrea.taverna@unimi.it

Abstract

We consider a large-scale unit commitment problem arising in medium-term simulation of energy networks, stemming from a joint project between the University of Milan and a major energy research centre in Italy. Optimal plans must be computed for a set of thermal and hydroelectric power plants, located in one or more countries, over a time horizon spanning from a few months to one year, with a hour-by-hour resolution. We propose a mixed-integer linear programming model for the problem. Since the complexity of this unit commitment problem and the size of real-world instances make it impractical to directly optimise this model using general purpose solvers, we devise ad-hoc heuristics and relaxations to obtain approximated solutions and quality estimations. We exploit an incremental approach: at first, a linear relaxation of an aggregated model is solved. Then, the model is disaggregated and the full linear relaxation is computed. Finally, a tighter linear relaxation of an extended formulation is obtained using column generation. At each stage, matheuristics are run to obtain good integer solutions. Experimental tests on real-world data reveal that accurate results can be obtained by our framework in affordable time, making it suitable for efficient scenario simulations.

1998 ACM Subject Classification G.1.6 Optimization

Keywords and phrases mathematical programming, unit commitment, power systems

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.63

1 Introduction

The Unit Commitment Problem (UCP) consists in finding the optimal production levels for plants with discrete activation patterns, i.e. plants that can be turned off and on. The body of literature on UCPs is huge and spans both theory and applications, as recent reviews like [1] and [2] report.

The most common objective is to minimise the global cost of production. Indeed, for global system simulations like ours, used, for instance, by power exchange authorities, the minimisation of global costs allows to maximise also the general welfare of the system. The UCP has traditionally been used to model thermal power plants schedules in power systems

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 63–75

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

on daily or weekly horizons to support operational decisions. However, recent applications have an increasing need for simulating energy networks on substantially longer periods, and including other types of plants such as hydroelectric, nuclear and waste-to-energy plants, as well as plants from renewable sources like wind, solar and biomass.

UCP models admit several variations that require different solving techniques. A distinctive characteristic is the length of the simulation horizon. Short-term models, from a few hours to weeks, are used to guide operational decisions: they are detailed but relatively small. These include complicating elements such as non-linear costs or resource consumption functions, to accurately represent thermal units. Medium and long-term models, that aim to support strategic decisions, are larger in size and span longer time horizons: they require more robust and efficient solution methods, but are less detailed.

UCP models typically include inter-temporal constraints for thermal plants, that bound or penalise the change in activity level of each unit between consecutive periods. For example, ramping constraints limit the change in production levels, minimum up and down time constraints prevent units from switching state too frequently, and start-up penalties model the costs that producers incur when their plants are switched on [3]. A large body of literature covers short-term UCPs. Among the approximated methods greedy algorithms [4] and meta-heuristic approaches [5, 6] have been proposed. These offer flexibility, but require fine-tuning to be effective, and provide no optimality guarantee. Mathematical programming is employed when better control on solutions quality is needed: short-term UCPs can be conveniently modelled as mixed-integer non-linear programs (MINLPs). Unfortunately, these MINLPs are in general too hard to be solved on real-sized instances. Only the very special case of single-unit has been successfully handled with exact algorithms [10]. Instead, Lagrangian relaxation schemes [7, 8, 9] are often used to obtain good approximations.

Linear models and mixed-integer linear programming (MILP) techniques have also been tried. Linear UCP approximations with a weekly horizon and more than 100 plants have been effectively solved via commercial MILP solvers [11], or embedded in rounding algorithms and used to solve real instances [12]. In [13] a branch-and-cut scheme allows to iteratively improve a model including a piecewise linear approximation of thermal unit costs.

Instances involving up to 100 thermal units and 200 hydroelectric plants on a weekly scenario were solved in this way. More recently, MINLP and MILP models have been coupled in a hybrid scheme [14].

Fewer publications deal with medium-term UCPs, spanning over months or years. Recent contributions include [15], in which a MILP model is devised for the simulation of the annual power production in Denmark. The authors combine a set of constructive heuristics, that compute approximated and potentially infeasible solutions, with improving methods, that compute feasible and better solutions starting from the heuristic ones. Computational results are reported on instances with 20 thermal plants: annual solutions are produced with computing times ranging between 5 and 10 minutes.

In this paper we face large-scale medium-term UCPs, to be solved on time periods ranging from a few months to the whole year. The aim is to support the simulation of the Italian energy market, where producers bid against one another to sell energy to buyers, i.e. private or public entities who provide energy to final consumers [16, 17]. Prices are then determined by the interaction between demand and supply. Cross-border flows are regulated: the pricing and the volumes exchanged are decided beforehand with a bidding mechanism according to a forecast of the future prices or demands in both systems. Indeed, mismatches in forecasting can cause discharges, blackouts, or simply the need to buy energy at very high prices, or sell it at negligible ones. Plant and system level features have to be accounted for. At plant-level,

both dispatchable (e.g. thermal units) and non-dispatchable power sources (e.g. renewable energy units) appear in the system. The latter cannot vary their production levels according to market players and thus their contribution to the supply is assumed to be known in advance through forecasting. Dispatchable power sources can be of either hydroelectric or thermal type. Hydroelectric plants, as other renewables, have negligible marginal production costs. Thermal plants are the most critical entities to model, having significant costs for fuel, periodic operations and maintenance, constraints and costs associated with the changes in the production levels and plants' states between consecutive periods. Finally, there are cap-pricing schemes that impose penalties for pollution, namely the production of CO_2 and NO_X gases, in each country and at the European level.

In Section 2 we introduce the model. In Section 3 we describe how to relax it in order to effectively obtain lower bounds on optimal solutions. In Section 4 we describe matheuristics and rounding procedures providing upper bounds. Finally, in Section 5 we report and discuss computational results on real-world instances, and in Section 6 we draw some conclusions.

2 Model

In the UCP variant we consider, plants exchange production through a power network. The network connects different zones, each hosting a set of hydroelectric and thermoelectric plants, by means of power links of limited capacity. Each zone has its own hourly demand to be satisfied by either plants in the zone or by import from nearby ones. Dispatching and production decisions have to be taken hour-by-hour. Hydroelectric plants have basins filled by lateral water inflow or pumping systems. Their production in each period linearly depends on the outbound flow from the basin and is assumed to be costless. Thermal plants need to be ignited and heated to be active. As such they have a binary activation state, fixed production costs, non-zero technical minimum and reduced flexibility, i.e. they have to maintain their state for a given amount of time. Production costs linearly depend on production levels and include pollution penalties. Some thermal plants have "double-shaft" technology; that is, they can switch between two working states, with one of them employing more power units and allowing for higher production levels. In our formulation thermal plants in each zone are partitioned in groups. Each of them is characterised by plants with identical marginal cost. Each group is further partitioned in families, having also the same technical minima, maxima and fixed costs.

Let T be the set of time periods. Let Z be the set of zones, $Y \subseteq Z$ denote the subset of zones from which energy can be exchanged with external systems, and $A \subseteq Z \times Z$ be the set of links between zones. Let

- H_z be the set of hydroelectric plants with reservoir in zone $z \in Z$, and $H = \bigcup_{z \in Z} H_z$,
- G_z be the set of groups of thermal plants for zone $z \in Z$, and $G = \bigcup_{z \in Z} G_z$,
- M_{zg} be the set of families of plants for zone $z \in Z$ and group $g \in G_z$, $M = \bigcup_{z \in Z, g \in G_z} M_{zg}$,
- $M^D_{zg} \subseteq M_{zg}$ be the subset of families of plants in zone $z \in Z$ and group $g \in G_z$ implementing double-shaft technology.

For each zone $z \in Z$, group $g \in G_z$ and family $m \in M_{zg}$, thermal plants are characterised by the following data:

- k_{zgm} and k^D_{zgm} (number of plants in family $m \in M_{zg}$ and $m \in M^D_{zg}$, resp.),
- c_{tzg} and e_{tzgm} (marginal and fixed production cost at time $t \in T$, resp.),
- p_{zgm} and P_{zgm} (minimum and maximum power produced by plants in family $m \in M_{zg}$, resp.),

- p^D_{zgm} and P^D_{zgm} (minimum and maximum power produced by plants in family $m \in M^D_{zg}$, resp., when in double-shaft mode),
- on_{zgm} and off_{zgm} (minimum time for which plants in family $m \in M_{zg}$ have to stay active, resp. inactive, once turned on, resp. off).

Let

$$T^{on}_{tzgm} = \{t' \in T : t \leq t' \leq \min(|T|, t + on_{zgm} - 1)\} \cup \{t' \in T : 1 \leq t' \leq (t + on_{zgm} - |T|)\}$$

be the set of periods in which a plant $m \in M_{zg}$ has to remain active if turned on at time $t \in T$, and

$$T^{off}_{tzgm} = \{t' \in T : t \leq t' \leq \min(|T|, t + off_{zgm} - 1)\} \cup \{t' \in T : 1 \leq t' \leq (t + off_{zgm} - |T|)\}$$

be the set of periods in which it has to remain inactive if turned off.

For each zone $z \in Z$, each hydroelectric plant $h \in H_z$ is characterised by the following data:

- p_h and P_h (minimum and maximum power produced, resp.),
- P^β_h (maximum pumping power),
- q_{zh} and Q_{zh} (volume available in the reservoir at the beginning of simulation, and required to be in the reservoir at the end of simulation, resp.),
- V_h (basin capacity),
- α_h and β_h (energy conversion and pumping efficiency coefficients),
- f_h and n_h (hourly maximum outflow and lateral inflow).

Furthermore, at each time period $t \in T$, let

- b_{tij} be the maximum energy transfer capacity of link $(i, j) \in A$,
- d_{tz} be the demand of zone $z \in Z$,
- E_t be the price of imported energy.

We introduce, for each period $t \in T$, zone $z \in Z$ and group $g \in G_z$, continuous variables x_{tzg} that represent the overall production level, and, for each family $m \in M_{zg}$, integer variables y_{tzgm}, y^D_{tzgm}, up_{tzgm}, dn_{tzgm}, up^D_{tzgm} and dn^D_{tzgm}, that represent the number of plants that are resp. active, active in double-shaft mode, switched on, switched off, entered and exited from double-shaft mode. We also consider, for each period $t \in T$, zone $z \in Z$ and plant $h \in H_z$, continuous variables $l_{tzh}, m_{tzh}, s_{tzh}, o_{tzh}$, that represent production level, pumping power, excess outbound flow from the basin, and reservoir volume, resp., for hydroelectric plants.

Finally, we assume the energy distribution network to have tree topology, since this is the case in Italy. Nevertheless our model can be extended to arbitrary structures. Continuous variables w_{tij} represent the amount of energy flowing through link $(i, j) \in A$ at time $t \in T$, and imp_{tz} and exc_{tz} the energy imported from external systems and the exceeding production in zone $z \in Y$ at time $t \in T$, respectively. As explained in the introduction, cross-border flows are regulated beforehand and unforeseen imports or exports are not expected. Therefore imp_{tz} and exc_{tz} variables are introduced only to detect issues in data forecasting or actual problems in the simulated system. Then E_t is meant to be set to a high value to minimise the use of imported energy and the exceeding production is assumed to be lost. Our UCP can be formulated as the following MILP.

$$\min \quad \phi = \sum_{\substack{t\in T, z\in Z,\\ g\in G_z}} c_{tzg}x_{tzg} + \sum_{\substack{t\in T, z\in Z,\\ g\in G_z, m\in M_{zg}}} e_{tzgm}y_{tzgm} + \sum_{t\in T, z\in Z} imp_{tz}E_t \tag{1a}$$

$$\text{s. t. } x_{tzg} \geq \sum_{m\in M_{zg}} p_{zgm}y_{tzgm} + \sum_{m\in M^D_{zg}} (p^D_{zgm} - p_{zgm})y^D_{tzgm} \ \forall t \in T, z \in Z, g \in G_z \tag{1b}$$

$$x_{tzg} \leq \sum_{m\in M_{zg}} P_{zgm}y_{tzgm} + \sum_{m\in M^D_{zg}} (P^D_{zgm} - P_{zgm})y^D_{tzgm} \ \forall t \in T, z \in Z, g \in G_z \tag{1c}$$

$$y^D_{tzgm} \leq y_{tzgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1d}$$

$$up_{tzgm} \geq y_{tzgm} - y_{(t-1)zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M_{zg} \tag{1e}$$

$$dn_{tzgm} \geq y_{(t-1)zgm} - y_{tzgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M_{zg} \tag{1f}$$

$$up^D_{tzgm} \geq y^D_{tzgm} - y^D_{(t-1)zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1g}$$

$$dn^D_{tzgm} \geq y^D_{(t-1)zgm} - y^D_{tzgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1h}$$

$$y_{tzgm} \geq \sum_{\tau\in T: t\in T^{on}_\tau} up_{\tau zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M_{zg} \tag{1i}$$

$$y_{tzgm} \leq k_{zgm} - \sum_{\tau\in T: t\in T^{off}_\tau} dn_{\tau zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M_{zg} \tag{1j}$$

$$y^D_{tzgm} \geq \sum_{\tau\in T: t\in T^{off}_\tau} up^D_{\tau zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1k}$$

$$y^D_{tzgm} \leq k^D_{zgm} - \sum_{\tau\in T: t\in T^{off}_\tau} dn^D_{\tau zgm} \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1l}$$

$$o_{1zh} = q_{zh} \ \forall z \in Z, h \in H_z \tag{1m}$$

$$o_{(|T|+1)zh} = Q_{zh} \ \forall z \in Z, h \in H_z \tag{1n}$$

$$o_{tzh} + n_{th} + \beta_h \cdot m_{tzh} = o_{(t+1)zh} + s_{tzh} + l_{tzh} \ \forall t \in T, z \in Z, h \in H_z \tag{1o}$$

$$\sum_{h\in H_z} \alpha_h \cdot l_{tzh} + \sum_{g\in G_z} x_{tzg} + \sum_{(i,z)\in A} w_{tiz} + \sum_{z\in Y} imp_{tz} \geq d_{tz} + \sum_{h\in H_z} m_{tzh} + \sum_{(z,j)\in A} w_{tzj} + \sum_{z\in Y} exc_{tz} \quad \forall t \in T, z \in Z \tag{1p}$$

$$y_{tzgm}, up_{tzgm}, dn_{tzgm} \in [0, k_{zgm}] \cap \mathbb{Z}^+_0 \ \forall t \in T, z \in Z, g \in G_z, m \in M_{zg} \tag{1q}$$

$$y^D_{tzgm}, up^D_{tzgm}, dn^D_{tzgm} \in [0, k^D_{zgm}] \cap \mathbb{Z}^+_0 \ \forall t \in T, z \in Z, g \in G_z, m \in M^D_{zg} \tag{1r}$$

$$w_{tij} \in [0, b_{ij}] \ \forall t \in T, (i,j) \in A \tag{1s}$$

$$s_{tzh} \in [0, f_h], o_{tzh} \in [0, V_h], l_{tzh} \in [p_h, P_h], m_{tzh} \in [0, P^\beta_h] \ \forall t \in T, z \in Z, h \in H_z \tag{1t}$$

$$imp_{tz} \geq 0, exc_{tz} \geq 0 \ \forall t \in T, z \in Y \tag{1u}$$

Constraints (1b) and (1c) impose that production level is 0 for inactive plants, and within production bounds for active ones. Constraints (1d) impose that only active plants can enter double-shaft mode. Constraints (1e)–(1h) enforce consistency between variables describing activation patterns. Constraints (1i)–(1l) impose that activation patterns respect minimum on and off times after switching. Finally (1m)–(1p) are flow conservation constraints ensuring energy balance between zones and consistency with thermal and hydroelectric productions inside each zone. The objective (1a) is to minimise the sum of production and additional energy import costs.

Due to the peculiarity of the UCP we consider, the resulting formulation (1) is significantly different than those previously proposed in the literature. A more detailed discussion on modelling issues is presented in [16] and [17].

3 Lower bounds

Preliminary experiments revealed that in large scale instances (a) even solving the continuous relaxation of formulation (1a)–(1u) (CR in the remainder) is computationally demanding, and (b) the bound obtained in this way has a non-negligible optimality gap. Therefore, we first propose to aggregate parts of the model, to obtain a relaxation that, although potentially weaker, can be solved more efficiently, coping with issue (a). Then we propose a decomposed model that has an exponential number of variables, but a reduced number of constraints. By optimising it through column generation we are able to obtain tighter bounds, coping with issue (b). These techniques are then meant to be used sequentially.

3.1 Aggregate Continuous Relaxation

For each $t \in T, z \in Z, g \in G_z$, let

$$\tilde{e}_{tzg} = \min_{m \in M_{zg}} \{e_{tzgm}\},$$
$$P^x_{zg} = \sum_{m \in M_{zg}} P_{zgm} k_{zgm} + \sum_{m \in M^D_{zg}} (P^D_{zgm} - P_{zgm}) k^D_{zgm}$$
$$\tilde{c}_{tzg} = \frac{\tilde{e}_{tzg}}{P^x_{zg}}.$$

We consider an aggregate continuous relaxation (ACR) given by the following model:

$$\min \quad \tilde{\phi} = \sum_{\substack{t \in T, z \in Z, \\ g \in G_z}} \tilde{c}_{tzg} x_{tzg} + \sum_{t \in T, z \in Z} imp_{tz} E_t \tag{2a}$$

$$\text{s.t. } 0 \le x_{tzg} \le P^x_{zg} \ \forall t \in T, z \in Z, g \in G_z \tag{2b}$$
$$(1m) - (1p)$$

that intuitively is a linear continuous model obtained by removing all integer variables, and by approximating the piecewise-linear cost function of each thermal group in each period, which may be non-continuous or non-differentiable, with a continuous linear lower-bound.

Model (2) can be shown to provide weaker bounds than CR, unless each group is composed by a single unit, in which case they coincide. On the other hand it can be solved more efficiently than CR as it is smaller and can be formulated as a network flow problem, for which well-known exact polynomial time algorithms can be used.

3.2 Decomposed model

For all $z \in Z, g \in G_z, m \in M_{zg}$ let

$$S_{zgm} = \{\left(\hat{y}^u_{tzgm}, \hat{y}^{uD}_{tzgm}, \hat{up}^u_{tzgm}, \hat{up}^{uD}_{tzgm}, \hat{dn}^u_{tzgm}, \hat{dn}^{uD}_{tzgm}\right)^{\intercal}_{t \in T} | (1d) - (1l), (1q) - (1r)\}$$

be the set of feasible activation patterns of thermal plants of a given family on the whole horizon. The following linear program represents the Dantzig-Wolfe reformulation [18] of the

continuous relaxation of model (1), where the feasible region defined by constraints (1d)–(1l), (1q)–(1r) has been replaced by the convex hull of its extreme integer points:

$$\min \quad \hat{\phi} = \sum_{\substack{t\in T, z\in Z,\\ g\in G_z}} c_{tzg}x_{tzg} + \sum_{\substack{z\in Z, g\in G_z,\\ m\in M_{zg}, u\in S_{zgm}}} \gamma^u_{zgm}(\sum_{t\in T} \hat{y}^u_{tzgm}e_{tzgm}) + \sum_{t\in T, z\in Z} imp_{tz}E_t \quad (3a)$$

$$\text{s.t. } x_{tzg} \geq \sum_{\substack{m\in M_{zg},\\ u\in S_{zgm}}} (\hat{y}^u_{tzgm}p_{zgm} + \hat{y}^{uD}_{tzgm}p^D_{zgm})\gamma^u_{zgm} \ \forall t \in T, z \in Z, g \in G_z \quad (3b)$$

$$x_{tzg} \leq \sum_{\substack{m\in M_{zg},\\ u\in S_{zgm}}} (\hat{y}^u_{tzgm}P_{zgm} + \hat{y}^{uD}_{tzgm}P^D_{zgm})\gamma^u_{zgm} \ \forall t \in T, z \in Z, g \in G_z \quad (3c)$$

$$\sum_{u\in S_{zgm}} \gamma^u_{zgm} = 1 \ \forall z \in Z, g \in G_z, m \in M_{zg} \quad (3d)$$

$$\gamma^u_{zgm} \in [0,1] \ \forall z \in Z, g \in G_z, m \in M_{zg}, u \in S_{zgm} \quad (3e)$$

$$(1m) - (1u)$$

The reformulation details are omitted. Indeed, without loss of optimisation potential, each set S_{zgm} can include only those patterns corresponding to extreme integer points. For each such a point $u \in S_{zgm}$, integer coefficients $\hat{y}^u_{tzgm}$ and $\hat{y}^{uD}_{tzgm}$ represent the number of plants that are active in normal and double-shaft mode, resp., in the corresponding pattern. Each variable γ^u_{zgm} is 1 if pattern $u \in S_{zgm}$ is fully selected, 0 if it is not selected at all. Fractional values are feasible: constraints (3d) and (3e) enforce that a linear convex combination of points in S_{zgm} is selected for each $z \in Z, g \in G_z, m \in M_{zg}$. Constraints (3c) and (3b) are the reformulated counterparts of constraints (1c) and (1b), resp..

▶ Proposition 1. The lower bound provided by (3) is at least as tight as that given by CR.

The proof follows immediately by the Dantzig-Wolfe decomposition principle [18].

Model (3) contains a combinatorial number of variables. In fact, neglecting double shaft and inter-temporal constraints, $|S_{zgm}| = k^{|T|}_{zgm}$. Therefore we optimise it by means of column generation techniques: we start with a restricted model where each set S_{zgm} is replaced by a subset $\overline{S}_{zgm}$ containing only patterns generated by heuristics. Then we iteratively solve the restricted model, obtain a vector of dual variables, and search for columns of negative reduced cost by solving the following *pricing problem*, for each $z \in Z$, $g \in G_z$ and $m \in M_{zg}$:

$$\min \pi^u_{zgm} = (\sum_{t\in T} \hat{y}^u_{tzgm}e_{tzgm}) - \eta_{zgm} +$$
$$- \sum_{t\in T}(\hat{y}^u_{tzgm}p_{zgm} + \hat{y}^{uD}_{tzgm}p^D_{zgm})\lambda_{zgm} - \sum_{t\in T}(\hat{y}^u_{tzgm}P_{zgm} + \hat{y}^{uD}_{tzgm}P^D_{zgm})\mu_{zgm} \quad (4a)$$

$$\text{s.t. } (\hat{y}^u_{tzgm}, \hat{y}^{uD}_{tzgm}, \hat{up}^u_{tzgm}, \hat{up}^{uD}_{tzgm}, \hat{dn}^u_{tzgm}, \hat{dn}^{uD}_{tzgm})_{t\in T} \in S_{zgm} \quad (4b)$$

where λ_{zgm}, μ_{zgm} and η_{zgm} are the dual variables associated with constraints (3b), (3c) and (3d), resp..

If any column of negative reduced cost is found, then the corresponding pattern is inserted in $\overline{S}_{zgm}$, and the process is iterated; otherwise, the optimal solution of the restricted model is optimal also for the full model, and therefore the process is halted.

We point out that each model (4), although being an integer linear program, asks to optimise a *single* family of plants, in a single zone and a single group. This makes it still manageable by general purpose solvers even for long time horizons.

4 Upper Bounds

In order to obtain good feasible integer solutions quickly, we designed several upper bounding heuristics [16, 17]. Two of them showed appealing results: an alternating matheuristic to be run after ACR and CR, described in subsection 4.1, and a rounding heuristic to be run at each column generation iteration, described in subsection 4.2.

4.1 Plan&Combine

As soon as an initial relaxation is computed, we run the following matheuristic, that we indicate as "Plan&Combine" (P&C). Intuitively, we first fix production levels and search for activation patterns of minimum cost for which the production levels are feasible (Plan). Then, we fix activation patterns and optimise production levels (Combine). We iterate Plan and Combine phases until no more changes are made in either phase.

Plan: for each $t \in T, z \in Z, g \in G_z$, let

$$\tilde{x}_{tzg} \in [0, \sum_{m \in M_{zg}} P_{zgm} k_{zgm} + \sum_{m \in M^D_{zg}} (P^D_{zgm} - P_{zgm}) k^D_{zgm}]$$

be a given set of feasible production levels.

We compute the minimum cost activation patterns allowing such production levels by solving the following integer linear program:

$$\min \phi_{Plan} = \sum_{\substack{t \in T, z \in Z, \\ g \in G_z, m \in M_{zg}}} e_{tzgm} y_{tzgm} \tag{5a}$$

$$\text{s.t. } \tilde{x}_{tzg} \leq \sum_{m \in M_{zg}} P_{zgm} |M_{zg}| + \sum_{m \in M^D_{zg}} (P^D_{zgm} - P_{zgm}) |M^D_{zg}| \quad \forall t \in T, z \in Z, g \in G_z \tag{5b}$$

$$(1d) - (1l)$$

It is easy to note that (5) decomposes in one independent subproblem for each $t \in T, z \in Z, g \in G_z$, making it well solvable with general purpose solvers.

Combine: for each $z \in Z, g \in G_z, m \in M_{zg}$, let

$$(\tilde{y}_{tzgm} \dots \tilde{y}^D_{tzgm})^\intercal_{t \in T} \in S_{zgm}$$

be a feasible activation pattern. We compute the minimum cost production respecting minimum and maximum power levels by solving the following linear program:

$$\min \phi_C = \sum_{\substack{t \in T, z \in Z, \\ g \in G_z}} c_{tzg} x_{tzg} + \sum_{\substack{z \in Z, g \in G_z, \\ m \in M_{zg}}} e_{tzgm} \tilde{y}_{tzgm} + \sum_{t \in T, z \in Z} imp_{tz} E_t \tag{6a}$$

$$\text{s.t. } x_{tzg} \leq \sum_{m \in M_{zg}} P_{zgm} \tilde{y}_{tzgm} + \sum_{m \in M^D_{zg}} (P^D_{zgm} - P_{zgm}) \tilde{y}^D_{tzgm} \ \forall t \in T, z \in Z, g \in G_z \tag{6b}$$

$$x_{tzg} \geq \sum_{m \in M_{zg}} p_{zgm} \tilde{y}_{tzgm} + \sum_{m \in M^D_{zg}} (p^D_{zgm} - p_{zgm}) \tilde{y}^D_{tzgm} \ \forall t \in T, z \in Z, g \in G_z \tag{6c}$$

$$(1d) - (1l)$$

that is comparable in complexity to ACR (2).

The P&C heuristic works as follows: for each $i \in \mathbb{N}_0^+$

1. let $\tilde{\mathbf{x}}^i$ be a feasible vector of production levels.
2. (plan) solve mod.(5) yielding activation patterns $\tilde{\boldsymbol{y}}^i$
3. (combine) solve mod.(6) yielding a feasible solution U^{i+1} for the original model (1). Let $\tilde{\boldsymbol{x}}^{i+1}$ be the corresponding production levels for thermal plants
4. if $\tilde{\boldsymbol{x}}^i \neq \tilde{\boldsymbol{x}}^{i+1}$ then $i := i + 1$, go to step 2. Otherwise stop.

It is worth noting that a solution given by a 'plan' step is always feasible for the subsequent 'combine' step, and a solution given by a 'combine' step is always feasible for the 'plan' step of the subsequent iteration. Hence the solutions produced never worsen during the iterations of P&C. It follows also that loops may occur only between solutions with the same value. Therefore, the convergence of P&C is guaranteed by considering a lexicographic order between solutions as a secondary objective function.

4.2 Column Generation rounding

At each column generation iteration we search for good integer solutions with the following Rounding Heuristic (RH). After pricing, we consider model (3): for each $z \in Z$, $g \in G_z$ and $m \in M_{zg}$, we fix to one the γ^u_{zgm} variable of highest fractional value, and we fix to zero all the remaining variables. Ties are broken according to the lexicographic order. In this way, no more integer variables are left free, and model (3) becomes a linear program: by optimising it using a suitable algorithm, like dual simplex, we obtain a full UCP solution.

5 Implementation and Results

We combined our algorithms to produce Upper Bounds (UB) and Lower Bounds (LB) on the optimal solution value with the following approach:

1. Solve ACR (obtain LB); run P&C starting from ACR optimal x_{tzg} values (obtain UB).
2. If (UB = LB) then stop (optimality is proved).
3. Solve CR (update LB); run P&C starting from CR optimal x_{tzg} values (update UB).
4. If (UB = LB) then stop (optimality is proved).
5. Populate sets $\bar{S}_{zgm}$ with solutions from steps 1 and 3, and from a pricing round using as dual values those corresponding to constraints (1c) and (1b) in the CR solution.
6. Run column generation until convergence, using RH at each iteration, updating UB.
7. Output the best UB found as final UB, and the final solution of model (3) as LB.

That is, we incrementally compute tighter bounds at the expense of higher computing efforts, stopping as soon as upper and lower bounds match.

We implemented our algorithms in AMPL [19], using IBM ILOG CPLEX 12.4 for solving both MILPs and LPs. CPLEX network simplex algorithm was used to solve ACR and Plan instances, while the barrier algorithm was selected for column generation LPs, including those in RH. P&C was stopped as soon as no improvements in the solution values were obtained, as in preliminary experiments no further improvement occurred afterwards. In each test, column generation was stopped after a time limit of $3600s$.

We performed a set of experiments on a notebook equipped with an Intel Core 2 Duo 1.2GHz processor, 4GB of RAM and running a Linux Operating System. As a benchmark, we used real data collected by RSE S.p.A.. They refer to the Italian energy market, and consist of $|Z| = 7$ zones in a tree network, three of which connected to external markets, 148 thermal plants partitioned in $|G| = 98$ groups and $|M| = 103$ families, and $|H| = 34$ groups of hydroelectric plants. Thermal plants are split in three parts, the first including 68 plants having minimum on and off times of 12 and 6 hours, resp., the second including 48

Table 1 CPLEX performance on Italian Energy Market data.

Size	Id	Size after presolving				CPU time (s)	gap %
		Constraints	Continuous var.	Integral var.	Binary var.		
1 month	1	152,541	110,184	13,870	67,158	-	33.34
	2	152,360	110,050	13,853	67,104	-	19.52
	3	151,649	109,485	13,750	66,852	3396	0.0
	4	151,578	109,451	13,740	66,836	1452	0.0
	5	151,624	109,477	13,753	66,831	842	0.0
	6	152,406	110,075	13,864	67,112	-	41.55
	7	152,266	109,956	13,855	67,062	-	< 0.01
	8	152,492	110,149	13,861	67,150	-	45.22
	9	152,147	109,879	13,819	67,013	-	0.11
	10	152,218	109,932	13,829	67,043	-	18.29
	11	152,311	109,992	13,838	67,093	2454	0.0
	12	151,774	109,591	13,813	66,908	-	19.65
2 months	1	304,887	220,249	27,721	134,246	-	31.42
	2	302,960	218,771	27,478	133,544	-	48.63
	3	304,033	219,579	27,616	133,931	-	< 0.01
	4	304,776	220,149	27,725	134,199	-	36.53
	5	304,423	219,867	27,646	134,080	-	46.56
	6	303,740	219,332	27,595	133,864	-	40.79
3 months	1	456,617	329,831	41,493	201,113	-	47.41
	2	455,838	329,210	41,386	200,855	-	59.97
	3	457,109	330,187	41,563	201,281	-	57.96
	4	456,073	329,383	41,460	200,918	-	57.11
6 months	1	912,782	659,305	82,904	402,076	-	60.74
	2	913,597	659,919	83,019	402,375	-	60.15
12 months	1	1,826,563	1,319,378	165,951	804,489	-	-

plants having minimum on and off times of 60 and 20 hours, resp., and the third including 32 plants with no constraints on minimum on and off times.

Demand data are given, and planning decisions required, for the full year with a hour-by-hour resolution, that is considering $T = 8760$ time slots. Besides testing our algorithms on the full 12 months horizon, we extracted three sets of instances corresponding to single months (12 instances, $T = 730$), pairs of consecutive months (6 instances, $T = 1460$), quarters (3 instances, $T = 2190$) and semesters (2 instances, $T = 4380$). The price of imported energy E_t was set to a very high value: our algorithms were always able to find solutions requiring neither energy import nor excess.

First, as a term of comparison, we performed a set of test by running the CPLEX MILP solver with default settings using model (1). The corresponding results are reported in Table 1 whose columns contain, in turn, instance size and reference, number of constraints, continuous, integer and binary variables after presolving, CPU time spent in optimising (or dash when a time limit of 4800s was hit), optimality gap at the end of computation. It can be noticed that such an approach leaves in general very large gaps even for small instances; CPLEX is able to close such a gap on four cases only, but the required computing time is very high. This preliminary check stresses the need for more computationally effective methods.

Then, we ran a set of tests with our incremental approach. In order to highlight the behaviour of each step of our method, and the relative impact of ACR and CR, we measured also intermediate bounds. Our results are reported in Table 2, which is composed by four

Table 2 Computational results of the incremental approach on Italian energy market data.

Size	Id	Continuous Lower Bounds		Plan&Combine				P&C + column generation			
		ACR	CR	after ACR		after CR		ACR init		CR init	
		Time (s)	Time (s)	Time (s)	Gap %	Time (s)	Gap %	Iter.	Gap %	Iter.	Gap %
1 month	1	7	30	46	3.55	70	0.38	30	1.44	42	0.22
	2	6	25	38	3.02	64	0.33	34	0.59	42	0.17
	3	7	25	39	2.89	65	0.30	42	0.41	50	0.17
	4	7	19	43	3.26	70	0.36	40	0.64	30	0.35
	5	6	22	39	3.07	61	0.32	38	1.44	44	0.13
	6	6	25	38	3.17	64	0.30	36	0.31	48	0.14
	7	7	26	45	3.10	68	0.29	34	0.54	41	0.16
	8	7	39	42	4.55	87	0.34	32	1.86	40	0.34
	9	8	33	42	3.20	71	0.29	36	0.52	37	0.15
	10	7	31	40	3.13	71	0.35	38	0.72	39	0.17
	11	7	27	44	3.10	67	0.32	39	0.45	44	0.15
	12	7	25	51	3.55	75	0.30	48	0.36	49	0.13
2 months	1	13	86	99	3.23	173	0.35	18	0.79	20	0.19
	2	14	84	81	2.94	168	0.33	17	0.42	20	0.12
	3	15	98	107	3.01	188	0.32	14	0.63	16	0.16
	4	18	69	97	3.57	171	0.32	18	0.75	19	0.18
	5	13	105	79	3.07	192	0.29	18	0.44	18	0.19
	6	13	105	80	3.00	194	0.31	19	0.67	20	0.15
3 months	1	23	178	153	3.09	321	0.34	11	0.76	12	0.23
	2	25	155	153	3.10	319	0.35	12	1.02	13	0.16
	3	25	112	176	3.39	271	0.30	12	0.64	11	0.28
	4	23	132	228	2.91	266	0.31	11	0.65	13	0.16
6 months	1	53	396	395	3.37	910	0.34	5	0.63	5	0.27
	2	59	303	379	3.44	891	0.33	6	1.29	4	0.22
12 months	1	125	1195	1301	3.50	-	-	-	-	-	-

blocks. The first block contains the instance size and reference. The second one contains the CPU time needed to compute ACR and CR. The third one refers to the P&C heuristics, and consists of two sub-blocks, reporting the cumulative CPU time needed to run both ACR and P&C (resp. CR and P&C) and the optimality gap $(UB - LB)/LB$ obtained, where UB is the value produced by P&C and LB is that given by ACR (resp. CR). The final block refers to the column generation process, and also consists of two sub-blocks, reporting the number of column generation iterations performed within the time limit and the optimality gap reached, when the sets $\bar{S}_{zgm}$ are initially populated with heuristic solutions obtained by running P&C after either ACR or CR.

For the 12 months full instance, the last six columns are marked with a dash, as we encountered out-of-memory problems while running P&C after CR. Therefore also the column generation process could not be started. As can be noticed by looking at columns in the second block, ACR can be computed five to ten times faster than CR. Results in the third block show that the subsequent effort for computing P&C is instead similar, but P&C on CR solutions produces much better approximations, always reaching an optimality gap lower than 0.4%. Finally, column generation is able to reduce the optimality gap, consistently reaching values below 1%, even if very few iterations are made. Computing times tend to increase slowly as the size of the instance increases, while the optimality gaps remain stable. A closer look at the computational details of our simulations reveal that Pricing and Plan models are easy to solve: CPLEX is able to solve most of the instances via presolving, and in any case is able to prove optimality at the root node of the branch-and-bound tree by performing a few LP iterations. As an overall assessment, best quality results are obtained with column generation, when the RMP is initialised with P&C using CR solutions. However,

such an approach is not viable on the full-scale instance. Instead, P&C using ACR solutions shows to offer a good trade off between solutions quality and computational scalability.

Due to the peculiar features of our problem, no direct comparison with methods from the literature is possible. However, with respect to similar applications like [15] we were able to tackle instances (a) involving 7 times more thermal plants and including (b) minimum up/-down constraints, (c) double shaft operating modes, (d) hydroelectric plants and (e) requiring energy transfers among zones, with a comparable computing effort and solutions quality. Moreover, the algorithms proposed in [15] require the fine-tuning of several parameters, while ours are almost parameter-free.

6 Conclusions

We faced a large-scale medium-term UCP arising in practice, introducing both compact and extended MILP models. We designed an incremental approach, computing lower bounds of increasing complexity and accuracy, and upper bounds exploiting the corresponding relaxations. We performed experiments on instances spanning a time horizon of up to one year. In all our tests, solutions within a few percentage points from optimality can be found very early in the incremental optimisation process. On instances with a time horizon up to six months and hour-by-hour resolution, our incremental approach reaches last stage, and provides solutions that, for practical purposes, are provably within negligible distance from optimality.

References

1 B. Saravanan, S. Das, S. Sikri, and D.P. Kothari. *A solution to the unit commitment problem – a review*. Frontiers in Energy, 7(2): 223–236, 2013.

2 R. Mallipeddi, and P.N. Suganthan. *Unit commitment - A survey and comparison of conventional and nature inspired algorithms*. International Journal of Bio-Inspired Computation, 6 (2):71–90, 2014.

3 M. Peik-herfeh, H. Seifi, and M.K. Sheikh-El-Eslami. *Two-stage approach for optimal dispatch of distributed energy resources in distribution networks considering virtual power plant concept*. European Transactions on Electrical Power, in press, 2014.

4 Y. Tingfang and T. O. Ting. *Methodological priority list for unit commitment problem*. Proceedings of the 2008 International Conference on Computer Science and Software Engineering, 1:176–179, Washington DC, USA, 2008.

5 N. P. Padhy. *Unit commitment – a bibliographical survey*. IEEE Trans. on Power Systems, 19(2):1196–1205, 2004.

6 S. Salam. *Unit commitment solution methods*. World Academy of Science, Engineering and Technology, 11, 2007.

7 A. Belloni, A. L. D. S. Lima, M. E. P. Maceira, and C. A. Sagastizabal. *Bundle relaxation and primal recovery in unit problems. The Brazilian case.*. Annals of Operations Research, 120(1):21–44, 2003.

8 A. Borghetti, A. Frangioni, F. Lacalandra, and C. A. Nucci. *Lagrangian heuristics based on disaggregated bundle methods for hydrothermal unit commitment*. IEEE Trans. on Power Systems, 18(1):313–323, 2003.

9 E. C. Finardi and E. L. Da Silva. *Solving the unit commitment problem of hydropower plants via lagrangian relaxation and sequential quadratic programming*. IEEE Trans. on Power Systems, 21(2):83–844, 2006.

10 A. Frangioni. *Solving nonlinear single-unit commitment problems with ramping constraints*. Operations Research, 54:775, 2006.

11 G. W. Chang, Y. D. Tsai, C. Y. Lai, and J. S. Chung. *A practical mixed integer linear programming based approach for unit commitment*. IEEE Power Engineering Society General Meeting, 1:221–225, 2004.

12 G. Migliavacca, A. Formaro, and A. Zani *Guida all'uso dell'interfaccia utente del simulatore MTSIM ed alla metodologia di creazione della base dati*, RSE technical report n. 08003692, 2008.

13 A. Frangioni, C. Gentile, and F. Lacalandra, *Tighter Approximated MILP Formulations for Unit Commitment Problems*. IEEE Trans. on Power Systems, 24(1): 105–113, 2009.

14 A. Frangioni, M. De Santis, and C. Gentile. *A new hybrid lagrangian-milp approaches for unit commitment problems*. Contribution at 26th European Conference on Operations Research, Rome, Italy, 2013.

15 N. H. Kjeldsen, and M. Chiarandini. *Heuristic solutions to the long-term unit commitment problem with cogeneration plants*. Computers and Operations Research, 39(2):269–282, 2012

16 A. Ceselli, and G. Righini. *Modelli ed algoritmi per problemi di ottimizzazione con variabili miste intere (Unit Commitment)*, Technical Report (in italian), RSE, Segrate, Italy, 2013.

17 A. Taverna. *Analisi e sperimentazione di algoritmi di programmazione intera per la simulazione annuale del sistema elettrico italiano (Unit Commitment Problem)*, Master's thesis, Dipartimento di Informatica, Università degli Studi di Milano, 2011.

18 G. Dantzig, and P. Wolfe. *Decomposition Principle for Linear Programs*, Operations Research 8(1), 1960.

19 R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*, Pacific Grove, CA: Brooks/Cole – Thomson Learning, 2003.

A Model-Based Heuristic to the Min Max K-Arc Routing for Connectivity Problem

Vahid Akbari and Sibel Salman

College of Engineering, Koc University
Istanbul, Turkey
{vakbarighadkolaei, ssalman}@ku.edu.tr

Abstract

We consider the post-disaster road clearing problem with the goal of restoring network connectivity in shortest time. Given a set of blocked edges in the road network, teams positioned at depot nodes are dispatched to open a subset of them that reconnects the network. After a team finishes working on an edge, others can traverse it. The problem is to find coordinated routes for the teams. We generate a feasible solution using a constructive heuristic algorithm after solving a relaxed mixed integer program. In almost 70 percent of the instances generated both randomly and from Istanbul data, the relaxation solution turned out to be feasible, i.e. optimal for the original problem.

1998 ACM Subject Classification G.1.6 Optimization, G.2.2 Graph Theory, G.2.3 Applications

Keywords and phrases Arc Routing Problem, Mixed Integer Programming, Heuristic, Network Connectivity, Road Clearance

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.76

1 Introduction

Arc routing problems have attracted the interest of researchers and have many application areas such as snow plowing, street sweeping, garbage collection, mail delivery, school bus routing and meter reading (see [8]). The problem addressed in this paper falls into the class of arc routing problems, but also contains network design and scheduling aspects. The main motivation of this section is to give an overview of arc routing problems and to introduce some problems which are related to our study.

In the *Chinese Postman Problem* (CPP), given a graph $G = (V, E)$, the problem is to determine a minimum cost closed walk, traversing each edge of G at least once. CPP can be solved in polynomial time if it is defined on an undirected or directed Network [7]. It can also be solved in polynomial time if it is defined on an mixed and even network [7], or on windy and Eulerian networks [11]. If there is a fleet of identical vehicles, say K vehicles, then the problem of finding K tours such that all the edges are covered with minimum total cost is called *K-CPP* (see [4] and [10]).

When only a subset of the edges are required to be traversed, the problem is called *Rural Postman Problem* (RPP). Eiselt et al. [8] thoroughly review studies on *RPP*. This class of problems are defined on a graph or on a multi-graph $G = (V, A)$, where V is the vertex set, A is the arc/edge set, and a non-negative cost matrix is associated with A. The graph may be directed, undirected or both (mixed). If there is a fleet of K vehicles, then the problem of covering the required edges with K tours that minimize the total cost is called *K-RPP*. In particular, *Min-Max K-vehicle Windy Rural Postman Problem* [5] is closer to the problem defined in our study as it minimizes the maximum tour cost.

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 76–88

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Given a connected graph $G = (V, E)$ and required subsets $E_R \subseteq E$ and $V_R \subseteq V$, *General Routing Problem* (GRP) is to find a minimum cost closed walk traversing the edges in E_R and visiting the vertices in V_R at least once [9]. A large number of well-known arc routing and vehicle routing problems are special cases of the *GRP*. For instance, When $V_R = \emptyset$ and $E_R = R$ we obtain *Rural Postman Problem*. In addition, if $E_R = E$, we have the *Chinese Postman Problem*. On the other hand, if $E_R = \emptyset$, the problem reduces to *Travelling Salesman Problem*. Corberan et al. [6] provide a branch and cut algorithm for the Windy GRP that solves optimally RPP instances with around 1000 nodes and 3000 edges.

The experience of past earthquakes reveals that the roadway elements are quite vulnerable, while the damages can seriously affect the transportation of products and people. A notable recent case of these earthquakes happened in Japan in 2011. The 2011 earthquake off the Pacific coast of Tohoku was a magnitude 9.0 (Mw) undersea megathrust earthquake off the coast of Japan that occurred at 14:46 JST (05:46 UTC) on Friday 11 March 2011. As a result of this earthquake Japan's transport network suffered severe disruptions. Almost 4000 road segments, 78 bridges and 29 railway locations were reported to be damaged. Accumulated debris in the downtown of Kamaishi City, Iwate Prefecture, and a damaged arterial (national highway 45) virtually isolated the community from rescue efforts and about 76 percent of the highways in the area were closed due to damage. These type of high-impact incidents cause the network to be disconnected due to blocked roads, impeding accessibility to hospitals and critical supply and shelter locations. In the immediate disaster response phase, to facilitate emergency transportation, a critical subset of the blocked roads should be cleared or opened to restore network connectivity in shortest time. Different from studies which are looking for tours including required edges, our main concern is connectivity of the network. In addition, the set of required arcs are not known in advance, and there is no requirement for the walks to be closed.

Recently, several studies focused on upgrading a road network or improving accessibility after a disastrous situation (e.g., [12] and [1]). To the best of our knowledge, the restoration of the roads after a disaster by routing a fleet of K vehicles in order to ensure connectivity of a network has not been addressed in the literature. In this paper, we define a new network optimization problem to address this topic. Since the problem combines arc routing and network design elements, it is called *Arc Routing for Connectivity Problem (ARCP)*, and since we are considering the case with K vehicles, we call it *K-ARCP*. The case with a single vehicle was defined and studied by Asaly and Salman [3], where a mixed integer programming (MIP) model was developed and applied to the case of Istanbul.

This paper is organized as follows: Next section gives a complete description of *K-ARC*. In the third section, we present the relaxed MIP model developed to solve *K-ARC*. We refer to this model as *R-MIP*. The fourth section gives an algorithm to extract the walk of every vehicle from the solution of *R-MIP*. In Section 5, we discuss the *feasibility algorithm* and how we make the solution of *R-MIP* feasible for *K-ARC*. Section 6 is devoted to data generation and results. We close with some final remarks in Section 7.

2 Problem Definition

We model our road network as an undirected connected graph $G = (V, A)$. There is a cost (time) c_{ij} associated with traversing all edges $(i, j) \in A : c_{ij} = c_{ji}$ and $c_{ij} > 0$. After the disaster, a set of edges, B, will be blocked and removal of them from the graph G, will make it disconnected. Traversing a blocked edge is not possible unless it is opened. The opening operation may involve clearing of the debris or repairing damaged segments. We

represent the associated extra opening cost (time) of a blocked edge $(i, j) \in B$ as b_{ij}, where $\forall (i, j) \in B, b_{ij} = b_{ji}$ and $b_{ij} > 0$. This additional time or cost incurs when a blocked edge is traversed for the first time. We assume that the edge opening times will be estimated by collecting post-disaster information on road conditions.

According to $(i, j) \in A : c_{ij} = c_{ji}$ and $\forall (i, j) \in B, b_{ij} = b_{ji}$, roads can be used in both directions in the disaster response phase. Since our model does not rely on these assumptions and it can handle the non-asymmetric case as well, these presumptions does not jeopardize the generality of the model.

Teams consisting of required equipment, machinery and personnel should be mobilized to clear the roads in shortest time. We refer to these teams as vehicles from now on. K vehicles are initially positioned at specified vertices. After the catastrophe, they complete a walk by working on the blocked edges assigned to them. We refer to the initial position of a vehicle as its *depot*. Note that there might be more than one vehicle positioned in a particular depot initially.

The edges that are not blocked form a graph $G_B = (V, A/B)$ with $|Q|$ components, where Q is the set of disconnected components. (Each Q is a connected sub-graph.) We are looking for a subset of blocked edges, R, such that $G_R = (V, A/B \cap R)$ is connected. In fact every subset R, is included of particular required edges which unblocking them, guarantee the connectivity of G_R. The solution identifies R and constructs a walk for each vehicle that starts at its depot. The objective is to minimize the maximum cost (time) walk among the vehicles.

Without loss of optimality, we can assume that if the walk of two vehicles includes the same blocked edge, the one that arrives first to the edge will unblock it. If a vehicle arrives at a node incident to a blocked edge while another vehicle is opening it, then it has to wait until the edge is unblocked.

Let us represent the walk of vehicle k by W_k. The total time of W_k consists of: 1) time of traversing the edges, $C(W_k)$; 2) time of road clearance, $B(W_k)$; and 3) waiting time, $A(W_k)$. The total walk time for vehicle k is calculated as:

$$T(W_k) = C(W_k) + B(W_k) + A(W_k) \tag{1}$$

The objective is to minimize the maximum value of $T(W_k)$ over $\forall k \in \{1, 2, ..., K\}$, where K shows the number of vehicles.

A simple example with two vehicles is shown in Figure 1. The blocked edges are shown with dashed lines. On the left, we see three components arising due to blocked edges. On the right, we see a feasible solution. The vehicles leave their depots, open several blocked edges and cross over some healthy edges in their walks. The walks end when the graph becomes connected.

3 A Mathematical Model for K-ARCP

Calculating the arrival time of the vehicles to the nodes complicates the model, since edges can be traversed multiple times. Therefore, we consider a relaxed problem (=*R-MIP*) such that the timing of the vehicles is ignored. After solving the *R-MIP*, we may encounter two situations: 1) We do not have timing problems in our relaxed mathematical model solution and the optimal solution of *R-MIP* is in fact the optimal solution of the *K-ARCP*. 2) If the relaxed model solution has timing conflicts, then we derive a feasible solution by modifying the assignment of opening tasks to the vehicles and inserting waiting time as necessary. The solution of the *R-MIP* gives a lower bound on the optimal solution to *K-ARCP*. An upper

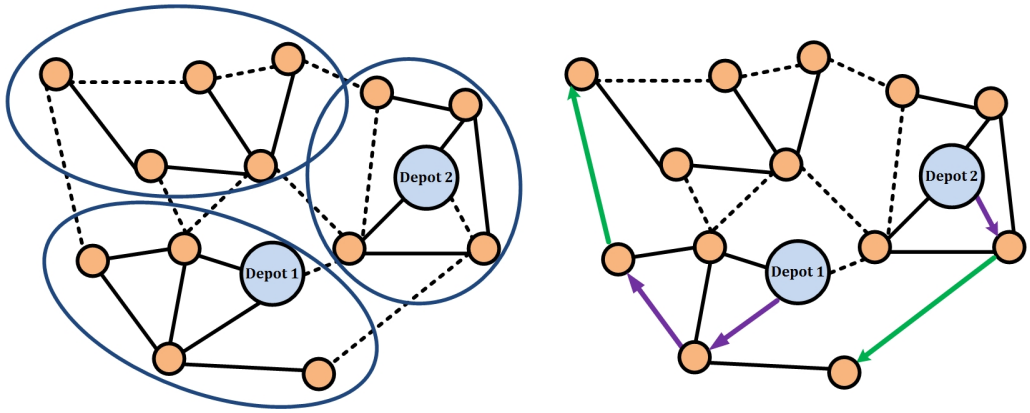


Figure 1 An illustrative example.

bound on the optimal solution of *K-ARCP* is obtained from the algorithm that constructs a feasible solution. In this way, we can derive an optimality gap for the feasible solution obtained.

Asaly [2] developed a model for the directed version of the same problem. All the time related constraints are also included as part of her model. However, since the number of variables and constraints are rapidly increasing with respect to the number of vehicles K and the size of the network, she could solve the model only for the single vehicle case, which does not require any waiting time. Asaly [2] proved that this problem is NP-hard even when a single vehicle exists. The NP-hardness of the symmetric *K-ARC* follows from this result.

The mixed integer programming (MIP) model for *K-ARCP* determines K open walks such that the disconnected components in the network are connected after unblocking a subset of the blocked edges. These walks start from depots and end in a dummy sink node, indexed as $n+1$. The model is formulated for the multi-depot and K vehicle case, where $K > 1$. The K walks altogether traverse a subset of the edges in B, denoted by R, so that the graph $G_R = (V, A/B \cap R)$ is connected. We define the decision variables and present the constraints next.

Decision Variables

$x_{ij}^k \in \{0,1\}$: indicates whether vehicle k traverses edge $(i,j) \in A$ from node i to j

$z_{ij}^k \in \{0,1\}$: indicates whether edge $(i,j) \in B$ is unblocked by vehicle k from node i to j

$f_{ij}^k \in \mathbb{N}_0$: flow of vehicle k on edge $(i,j) \in A$ from node i to j

$v_i^k \in \mathbb{N}_0$: number of times vehicle k visits node $i \in V$

Objective Function

Minimize y

subject to

$$\sum_{(i,j)\in A} c_{ij}x_{ij}^k + \sum_{(i,j)\in B} b_{ij}z_{ij}^k \leq y, \quad \forall k = 1,2,\ldots,K \tag{2}$$

Constraints. Let D denotes the set of depots and P_d shows the set of vehicles which are initially positioned in depot $d \in D$. Constraints (3),(4) and (5) are vehicle flow balance equations. Constraint (3) ensures that every vehicle leaves its depot and (4) and (5) enforces vehicles not to stop in any intermediate nodes. Constraint (6) forces every walk to end in the sink node. Each vehicle visits the sink node exactly once and does not get out. The latter case is satisfied by constraints (7).

$$\sum_{j \in V \cup \{(n+1)\}} (x^k_{dj} - x^k_{jd}) = 1, \quad \forall d \in D, \quad \forall k \in P_d \tag{3}$$

$$\sum_{j \in V \cup \{(n+1)\}} (x^k_{dj} - x^k_{jd}) = 0, \quad \forall d \in D, \quad \forall k \notin P_d \tag{4}$$

$$\sum_{j \in V \cup \{(n+1)\}} (x^k_{ij} - x^k_{ji}) = 0, \quad \forall k = 1, 2, \ldots, K, \quad \forall i \in V \backslash D \tag{5}$$

$$\sum_{j \in V} x^k_{j(n+1)} = 1, \quad \forall k = 1, 2, \ldots, K \tag{6}$$

$$x^k_{(n+1)i} = 0, \quad \forall i \in V, \quad \forall k = 1, 2, \ldots, K \tag{7}$$

The following sets of constraints establish a relation between z^k_{ij} and x^k_{ij} variables. Constraint (8) shows that if a blocked edge is opened it must be traversed at least once. Constraint (9) prevents traversing a blocked edge if it is not unblocked by any of the vehicles. Let us show the set of blocked edges in the cut-sets between components with C. When a vehicle leaves its depot, it may open all the blocked edges that are in C. We show this property with (10). Since our graph is undirected, it is enough to open a blocked edge in one direction, if it is selected to be opened. This is shown by (11).

$$x^k_{ij} \geq z^k_{ij}, \quad \forall k = 1, 2, \ldots, K, \quad \forall (i,j) \in B \tag{8}$$

$$x^k_{ij} \leq \sum_{\kappa=1}^{K} (z^\kappa_{ij} + z^\kappa_{ji}), \quad \forall k = 1, 2, \ldots, K, \quad \forall (i,j) \in B \tag{9}$$

$$\sum_{(i,j) \in C} z^k_{ij} \leq |(i,j) \in C| \sum_{j \in V : (d,j) \in A} x^k_{dj}, \quad \forall d \in D, \quad \forall k \in P_d \tag{10}$$

$$\sum_{\kappa=1}^{K} (z^\kappa_{ij} + z^\kappa_{ji}) \leq 1, \quad \forall (i,j) \in B \tag{11}$$

In order to ensure connectivity of the walks, we define flow variables f^k_{ij} for every vehicle and for each edge that it passes through. For depot vertices, the net flow into a depot vertex is the total number of visits to all vertices except the depot. For the other vertices, it is equal to the number of visits to the corresponding node. In other words, a vehicle leaves one unit of flow each time it visits a node. Constraints (15) ensure that walks end in sink node by sending one unit of flow to the sink node. (16) prevent backward flow from the sink node to any other node.

$$\sum_{j:(i,j)\in A,\{i,j\}\in V\cup\{(n+1)\}} (f_{ij}^k - f_{ji}^k) = -v_i^k,$$
$$\forall k = 1,2,\ldots,K, \quad \forall i \in V\cup\{(n+1)\}\setminus D \tag{12}$$

$$\sum_{j\in V\cup\{(n+1)\}} (f_{dj}^k - f_{jd}^k) = \sum_{i\in V\cup\{(n+1)\}\setminus\{d\}} v_i^k, \quad \forall k \in P_d, \quad \forall d \in D \tag{13}$$

$$\sum_{j:(d,j)\in A,\{i,j\}\in V\cup\{(n+1)\}} (f_{dj}^k - f_{jd}^k) = -v_d^k, \quad \forall d \in D, \quad \forall k \notin P_d \tag{14}$$

$$f_{(n+1)j}^k = 0, \quad \forall j \in V, \quad \forall k = 1,2,\ldots,K \tag{15}$$

$$\sum_{j\in V} f_{j(n+1)}^k = 1, \quad \forall k = 1,2,\ldots,K \tag{16}$$

Here, n shows the total number of nodes in our network. Constraints (17) do not allow flow on an edge unless it is traversed. Constraints (18) show that if an edge is traversed, then there must be a positive amount of flow passing through it.

$$f_{ij}^k \le (n-1)x_{ij}^k, \quad \forall k = 1,2,\ldots,K, \quad \forall (i,j) \in A, \quad \{i,j\} \in V\cup\{(n+1)\} \tag{17}$$

$$f_{ij}^k \ge x_{ij}^k, \quad \forall k = 1,2,\ldots,K, \quad \forall (i,j) \in A, \quad \{i,j\} \in V\cup\{(n+1)\} \tag{18}$$

A vertex is visited if and only if an edge entering that vertex is traversed, by constraint (19). If we cross a particular vertex more than one time, we should go to open another blocked edge that is in C. Constraint (20) enforces this.

$$\sum_{j:(i,j)\in A} x_{ji}^k = v_i^k, \quad \forall k = 1,2,\ldots,K, \quad \forall i \in V\cup\{(n+1)\} \tag{19}$$

$$v_i^k = \sum_{(i,j)\in C} z_{ij}^k, \quad \forall k = 1,2,\ldots,K, \quad \forall i \in V\cup\{(n+1)\} \tag{20}$$

Constraints (21) enforce connectivity of the graph. It is a sub-tour elimination constraint.

$$\sum_{\kappa=1}^{K} \sum_{(i,j)\in\delta(s)} z_{ij}^\kappa \ge 1, \quad \forall S \subset V, S \ne \emptyset \tag{21}$$

4 Walk Extraction Algorithm

After solving *R-MIP* we obtain values of x_{ij}^k and f_{ij}^k and v_i^k but we do not know about the order of edges to be visited by each vehicle. What we know is that whether a vehicle crosses an edge in a particular direction or not.

We define the *Walk Extraction Algorithm* for k^{th} vehicle as follows.

Walk Extraction Algorithm

- **step 1:** For the k^{th} vehicle if $\exists i \in V : v_i^k \ge 2$ go to step 3 and otherwise go to step 2.
- **step 2:** The decreasing order of f_{ij}^k values gives us the walk.
- **step 3:** In the graph $\Gamma = (V, \hat{A})$ such that V is the set of all vertices as before and $\hat{A} = \{(i,j) | x_{ij}^k = 1\}$; find the shortest path between P_d and the sink node.
- **step 4:** Using a *cycle detection* method, find all the loops in Γ.

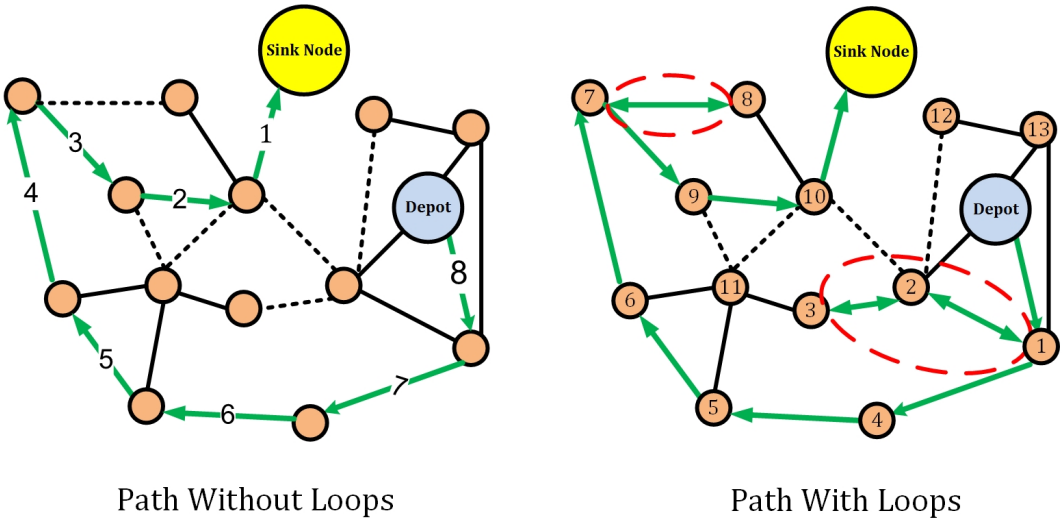


Figure 2 Samples of paths with and without loops.

- **step 5:** Go through nodes in the shortest path obtained in step 3 and if any of the nodes are in one of the loops derived from step 4; add the loop to the shortest path and remove the loop from the set of loops and then start step 5 again. If the size of loops set is 0, the path is complete.

There is an important property with the walks; they either contain loops or not. In the walk of the k^{th} vehicle a loop occurs if $\exists i \in V : v_i^k \geq 2$. It means that starting from node i, vehicle k visited a set of edges and came back to node i again. On the other hand, if in the k^{th} vehicle's walk $v_i^k \leq 1$ for all $i \in V$ there is no loop in it's walk. Step 1 determines whether the walk includes walks or not. Second step relation follows by the structure of flow variables. A sample of walk without loops is shown in the left side of Figure 2. Flow variable values on the edges are determining the path. In the right side network of Figure 2, a *cycle-included* walk is shown. In this case, the result of third step is: $Start \Rightarrow 1 \Rightarrow 4 \Rightarrow 5 \Rightarrow 6 \Rightarrow 7 \Rightarrow 9 \Rightarrow 10 \Rightarrow End$. For step 4, we used $simple - cycles(G)$ function from $NetworkX$ library in $Python$. For the given sample the loops are: $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1$ and $7 \Rightarrow 8 \Rightarrow 7$. Due to step 5 the output of the algorithm is as follows: $Start \Rightarrow 1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1 \Rightarrow 4 \Rightarrow 5 \Rightarrow 6 \Rightarrow 7 \Rightarrow 8 \Rightarrow 7 \Rightarrow 9 \Rightarrow 10 \Rightarrow End$.

▶ **Lemma 1** (Walk Extraction Algorithm). *Given the values of x_{ij}^k and f_{ij}^k and v_i^k $\forall k = 1, \ldots, K$, we can find the walk of each vehicle with the Walk Extraction Algorithm.*

Proof. The proof follows by the structure of the variables in the mathematical model and observations discussed in the algorithm about the properties of the walks. ◀

▶ **Lemma 2** (*R-MIP* Solution). *For $W_k, k = 1, \ldots, k$ showing the walk of the k^{th} vehicle and $R = B \cap W_k$, when R is added to G_B, we get a connected graph. Moreover, $T_R \leq Z^*_{K-ARCP}$ where T_R shows the optimal objective value of* R-MIP *and Z^*_{K-ARCP} denotes the optimal objective value of* K-ARCP.

Proof. The first part follows by the structure of the mathematical model and problem definition. Constraint 21 enforces to open at least a blocked edge between two separated components for every choice of subcomponents. This will result in finding $G_R = (V, A/B \cap R)$

Table 1 Order of edges visited by vehicles (1 to k).

Walk	Path of the walks	Finishing Time of the walks
W_1	$w_{1,1} \Rightarrow w_{1,2} \Rightarrow w_{1,3} \Rightarrow \ldots \Rightarrow w_{1,n(1)}$	$t_{1,1} \Rightarrow t_{1,2} \Rightarrow t_{1,3} \Rightarrow \ldots \Rightarrow t_{1,n(1)}$
W_2	$w_{2,1} \Rightarrow w_{2,2} \Rightarrow w_{2,3} \Rightarrow \ldots \Rightarrow w_{2,n(2)}$	$t_{2,1} \Rightarrow t_{2,2} \Rightarrow t_{2,3} \Rightarrow \ldots \Rightarrow t_{2,n(2)}$
.	.	.
.	.	.
.	.	.
W_k	$w_{k,1} \Rightarrow w_{k,2} \Rightarrow w_{k,3} \Rightarrow \ldots \Rightarrow w_{k,n(k)}$	$t_{k,1} \Rightarrow t_{k,2} \Rightarrow t_{k,3} \Rightarrow \ldots \Rightarrow t_{k,n(k)}$

Table 2 Order of edges from B visited by vehicles (1 to k).

Walk	Order of blocked edges	Finishing Time of the walks
W_1	$B_{1,1} \Rightarrow B_{1,2} \Rightarrow B_{1,3} \Rightarrow \ldots \Rightarrow B_{1,m(1)}$	$\tau_{1,1} \Rightarrow \tau_{1,2} \Rightarrow \tau_{1,3} \Rightarrow \ldots \Rightarrow \tau_{1,m(1)}$
W_2	$B_{2,1} \Rightarrow B_{2,2} \Rightarrow B_{2,3} \Rightarrow \ldots \Rightarrow B_{2,m(2)}$	$\tau_{2,1} \Rightarrow \tau_{2,2} \Rightarrow \tau_{2,3} \Rightarrow \ldots \Rightarrow \tau_{2,m(2)}$
.	.	.
.	.	.
.	.	.
W_k	$B_{k,1} \Rightarrow B_{k,2} \Rightarrow B_{k,3} \Rightarrow \ldots \Rightarrow B_{k,m(k)}$	$\tau_{k,1} \Rightarrow \tau_{k,2} \Rightarrow \tau_{k,3} \Rightarrow \ldots \Rightarrow \tau_{k,m(k)}$

such that G_R is connected. The second part follows by the fact that we are ignoring timing conflicts in *R-MIP*. In *R-MIP* we get a wider feasible region by ignoring timing variables and constraints, which will result in better optimal values. Disregarding time related elements will result in non feasible solutions in some cases. This infeasible cases occur when a vehicle is crossing a blocked edge since it is going to be unblocked by another vehicle. Since we are relaxing time constraints, the opener vehicle might unblock the blocked edge after the other vehicle crosses it in sense of time, which is not acceptable in *K-ARCP*. ◀

5 Feasibility Algorithm

Since *R-MIP* does not have time-related elements, our solution might not be feasible for *K-ARCP*. We can derive a feasible solution by modifying the solution obtained from *R-MIP*. We determine those edges with timing conflict and then we shift the time of the second vehicle that is crossing the blocked edge to derive a feasible solution.

As explained in the *Walk Extraction Algorithm* we can derive walks of each vehicle from the output of *R-MIP*. With the result obtained from the *Walk Extraction Algorithm* we can form Table 1. It includes the path of every vehicle and the corresponding time for crossing every edge in every path. $w_{i,j}$ shows j^{th} edge crossed by vehicle i and $t_{i,j}$ is the corresponding time with it. With the information in Table 1, $T_R = \max_k(t_{k,n(k)})$.

In Table 2, we only consider the edges in B that are traversed with different vehicles and the corresponding time of crossing them. Here $B_{i,j}$ shows j^{th} edge in B that has been crossed by vehicle i and $\tau_{i,j}$ is the time when traversing the edges is completed.

Note that an edge of set B may be repeated in the walk of a vehicle, or it may appears in the walks of more than one vehicles. During the algorithm, as we process the edges given in Table 2, we shift the finishing time of those edges that are effected and remove the processed edges from Table 2.

With Table 2 in hand we can write our feasibility algorithm as follows.

Feasibility Algorithm

- **Step 1:** Choose $\psi = min\{\tau_{i,j}\}$ over all $i = 1, ..., k, j = 1, ..., n(k)$.
  If $\psi = min\{\tau_{i,j}\} = max\{\tau_{i,j}\}$ paths are synchronized. Otherwise go to step 2.
- **Step 2:** Determine vehicle number $(= \kappa)$ and blocked edge $= (i, j)$ associated with ψ. (Go to one of the following cases considering z^{κ}_{ij} and z^{κ}_{ji} values:)
- **Case 1: If $z^{\kappa}_{ij} = 1$: "No Change Case"**
  "Apply following change:"
  Update Table 2 by removing all $B_{i,j}$s and $B_{j,i}$s and their corresponding time from it and go back to step 1.
- **Case 2: If $z^{\kappa}_{ij} = 0$ but $z^{\kappa}_{ji} = 1$: "Backward Opener Case"**
  "Apply following changes:"

  1. For $\forall \tau^{\kappa}_{\alpha,\beta} : \tau^{\kappa}_{i,j} \leq \tau^{\kappa}_{\alpha,\beta} \leq \tau^{\kappa}_{j,i}$, shift $\tau^{\kappa}_{\alpha,\beta}$ values as follows: $\tau^{\kappa}_{\alpha,\beta_{(new)}} = \tau^{\kappa}_{\alpha,\beta} + b_{ij}$ where b_{ij} is the extra cost associated with opening edges (i, j).
  2. If (i, j) or (j, i) is in the walk of vehicle η such that $\eta \neq \kappa$ and $\tau^{\eta}_{i,j} \leq \tau^{\kappa}_{i,j}$ or $\tau^{\eta}_{j,i} \leq \tau^{\kappa}_{i,j}$ for $\forall \tau^{\eta}_{\alpha,\beta} : \tau^{\kappa}_{i,j} \leq \tau^{\eta}_{\alpha,\beta}$, shift $\tau^{\eta}_{\alpha,\beta}$ values as following: $\tau^{\eta}_{\alpha,\beta_{(new)}} = \tau^{\eta}_{\alpha,\beta} + \tau^{\kappa}_{i,j} - \tau^{\eta}_{i,j} + c_{ij}$, where c_{ij} is the time or cost of crossing edge (i, j) after fixing it.
  3. Update Table 2 by removing all $B_{i,j}$s and $B_{j,i}$s and their corresponding time from it and go back to Step 1.

- **Case 3: If $z^{\kappa}_{ij} = 0$ and $z^{\kappa}_{ji} = 0$: "Swap opener vehicle Case"**
  Determine vehicle ρ such that $z^{\rho}_{ij} + z^{\rho}_{ji} = 1$ meaning vehicle ρ is opening edge (i, j). (For simplicity of notation let us assume vehicle ρ is opening edge (i, j) at time τ_{ρ}.)
  "Apply following changes:"

  1. (Making changes in the walk of vehicle ρ) $\forall \tau^{\rho}_{\alpha,\beta} : \tau_{\rho} \leq \tau^{\rho}_{\alpha,\beta}$, shift $\tau^{\rho}_{\alpha,\beta}$ values as following: $\tau^{\rho}_{\alpha,\beta_{(new)}} = \tau^{\rho}_{\alpha,\beta} - \tau_{\rho} + max\{\tau^{\kappa}_{i,j} + c_{ij}, \tau_{\rho} - b_{ij}\}$.
  2. (Making changes in the walk of vehicle κ) vehicle κ is now the opener of this blocked edge. So $\forall \tau^{\kappa}_{\alpha,\beta} : \tau^{\kappa}_{i,j} \leq \tau^{\kappa}_{\alpha,\beta}$, shift $\tau^{\kappa}_{\alpha,\beta}$ values as following : $\tau^{\kappa}_{\alpha,\beta_{(new)}} = \tau^{\kappa}_{\alpha,\beta} + b_{ij}$.
  3. If (i, j) or (j, i) is in the walk of vehicle η such that $\eta \neq \kappa \neq \rho$ and $\tau^{\eta}_{i,j} \leq \tau^{\kappa}_{i,j}$ or $\tau^{\eta}_{j,i} \leq \tau^{\kappa}_{i,j}$ for $\forall \tau^{\eta}_{\alpha,\beta} : \tau^{\kappa}_{i,j} \leq \tau^{\eta}_{\alpha,\beta}$, shift $\tau^{\eta}_{\alpha,\beta}$ values as following: $\tau^{\eta}_{\alpha,\beta_{(new)}} = \tau^{\eta}_{\alpha,\beta} + \tau^{\kappa}_{i,j} - \tau^{\eta}_{i,j} + c_{ij}$.
  4. Update Table 2 by removing all $B_{i,j}$s and $B_{j,i}$s and their corresponding time from it and go back to Step 1.

"No change" case: Since vehicle κ is opening edge (i, j) in its first visit there is no timing conflict in this case.

"Backward opener" case: In this case, same vehicle opens edge (i, j) in the (j, i) direction in its second pass over it. In the 2nd change of the "Backward opener" case we are considering that vehicle η could finish his walk on edge (i, j) at any possible time and particularly $\tau^{\eta}_{i,j}$ or $\tau^{\eta}_{j,i}$, but now the earliest time it can finish its walk across edge (i, j) is $\tau^{\kappa}_{i,j} + c_{ij}$.

"Swap opener vehicle" case: In this case, the vehicle κ is crossing edge (i, j) without unblocking it, since in the solution of *R-MIP*, ρ is the opener vehicle of edge (i, j), i.e, $z^{\rho}_{ij} + z^{\rho}_{ji} = 1$. Since ρ is opening this edge after vehicle κ crosses it, timing conflict happens. For the shifting procedure we should consider vehicle κ, ρ and all other vehicles k, such that $x^{k}_{ij} + x^{k}_{ji} \geq 1$ separately.

Table 3 Distance limit for graph generation.

Number of nodes	100	75	50	25
Distance	10	15	20	25

▶ **Lemma 3** (Feasibility Algorithm). *Given k walks through $G = (V, A)$; $w_{i,1} \Rightarrow w_{i,2} \Rightarrow w_{i,3} \Rightarrow \ldots \Rightarrow w_{i,n(i)}$ where $i \in \{1, 2, \ldots, k\}$, and their corresponding time ; $t_{i,1} \Rightarrow t_{i,2} \Rightarrow t_{i,3} \Rightarrow \ldots \Rightarrow t_{i,n(i)}$, by applying feasibility algorithm, we can modify the non-feasible walks to a feasible solution for* K-ARCP.

Proof. The proof follows by the properties of the algorithm. We keep track of the walks like a simulation system and assign required waiting times whenever it is necessary. With this procedure we prevent timing conflicts among vehicles on blocked edges. ◀

6 Results

In order to test the performance of *R-MIP* and the feasibility algorithm, we generated two types of data.

6.1 Randomly generated data set

We generated Euclidean random graphs with 25, 50, 75 and 100 nodes. First, we assigned random coordinates in a 100 × 100 plane to every node. Costs ($= c_{ij}$) on edges are equal to the Euclidean distances. The extra cost ($= b_{ij}$) on edge (i, j) is generated according to: $(b_{ij} = c_{ij} \times U(10, 30))$ where $U(10, 30)$ is uniform distribution between 10 and 30. In each case with different number of nodes, an edge (i, j) exists if distance between nodes i and j is lower than the distance limits given in Table 3. But, as we know, one of our primary assumptions is that our graph is connected in the beginning. In case the generated graph is not connected, we add some random edges between disconnected components to make it connected. Depots are also chosen randomly among all the nodes in the graph and we assumed every node has the potential to be a depot. According to the problem definition $G_B = (V, A/B)$ is a disconnected graph consisting of $|Q|$ separated components. We also impose $|Q| \geq K + 1$ in our instances, to increase the possibility of assigning at least one unblocking task to every vehicle. Edges are randomly added to the set B one by one. This procedure stops when $|Q| \geq K + 1$.

We had instances with 100, 75, 50 and 25 nodes. With 100 and 75 nodes, we tried 1, 2, 3 and 4 vehicles and for the cases with 50 and 25 nodes, we tried 1 and 2 vehicles. For every case, we had 5 different instances. For example, we generated 5 instances with the 75 nodes and 2 vehicles, and so on. In order to check the capabilities of our model, we tested 5 instances with 4 vehicles and 1000 nodes positioned in an $10,000 \times 10,000$ plane with edge existence distance of 10 units.

As we see in Table 4, in almost 70% of the instances we derived the optimal solution of *K-ARCP* by solving *R-MIP*. In the other cases we found good lower and upper bounds for the optimal solution of *K-ARCP*. When the number of vehicles gets higher, the possibility of occurrence of a timing conflict is higher and we get wider boundaries for our optimal solution. In the case with one vehicle, we can always get the optimal solution of *K-ARCP* with solving *R-MIP*, since timing conflict requires at least two vehicles.

In Table 4 and 5, the maximal optimality gap shows the maximum possible gap from optimal solution to either of the lower bound or upper bound solutions.

Table 4 Results with randomly generated data according to method 1.

Number of Nodes	Number of Vehicles	Maximal Optimality Gap	Exact Optimal Solution
1000	4	4.1 %	4 out of 5
100	1	0 %	5 out of 5
	2	2.6 %	4 out of 5
	3	3.2 %	3 out of 5
	4	8.4 %	2 out of 5
75	1	0%	5 out of 5
	2	5.2 %	2 out of 5
	3	5.8 %	2 out of 5
	4	5.0 %	2 out of 5
50	1	0%	4 out of 5
	2	2.4 %	5 out of 5
25	1	0 %	5 out of 5
	2	12.4 %	5 out of 5

We should mention that all of these instances were tested on an Intel® Xeon® E5-2643 0 CPU @ 3.3 GHz (two processors) with 32 GB of RAM device and except for the cases with 1000 nodes, all instances were solved in less than 1 minute. However, for the cases with 1000 nodes, all the instances were solved in at most 13 minutes. More than 80% of these run times were from solving *R-MIP*. Even with very large instances the *feasibility algorithm* takes less than a minute.

6.2 Istanbul data set

We used the network of Istanbul city given in [3]. Istanbul's network is modelled by 74 nodes and 316 edges. The edges are categorized into 3 different groups due to their proximity to the epicentre of the earthquake scenario: as high, medium and low risk edges. The probability to lose an edge after an earthquake, from low risk class is 0.3, and this probability is equal to 0.4 and 0.5 for medium and high risk edges, respectively. After each run we verified if the number of disconnected components, $|Q|$, is higher than the number of vehicles, K, or not. If $|Q| \geq K$ it means that our current number of vehicles is suitable for our problem, otherwise, we are keeping too many vehicles for our network. In the latter case, we do not solve the problem and we refer to them as NS in Table 5. In Table 5, exact optimal solution for every instance shows if the optimal solution of *R-MIP* and *K-ARC* are equal to each other or not. The fifth column in Table 5 gives the optimal objective value for every instance. This optimal value is the length of the longest walk among all the vehicles in hours. For Istanbul's network, we tested cases with 1 up to 4 vehicles. In each case, we generated 5 random problems. The results showed that keeping more than 3 vehicles for the probabilities assigned to different categories of edge risks is not logical, since all the 5 problems generated for this case had less than 5 components. Meanwhile, the probabilities that we assigned to losing edges after a disaster is quite pessimistic, which means 3 vehicles would be enough to support Istanbul's network in an expected earthquake.

For the case of 3 vehicles, in 2 out of 5 problems generated, we had 4 or more disconnected components and with the case of 2 vehicles, in 4 out of 5, our network was separated to more than 3 components. Table 5 shows all the results related to our instances. Table 5, shows that we obtained the optimal solution to *K-ARCP* by solving *R-MIP* in almost 60% of the

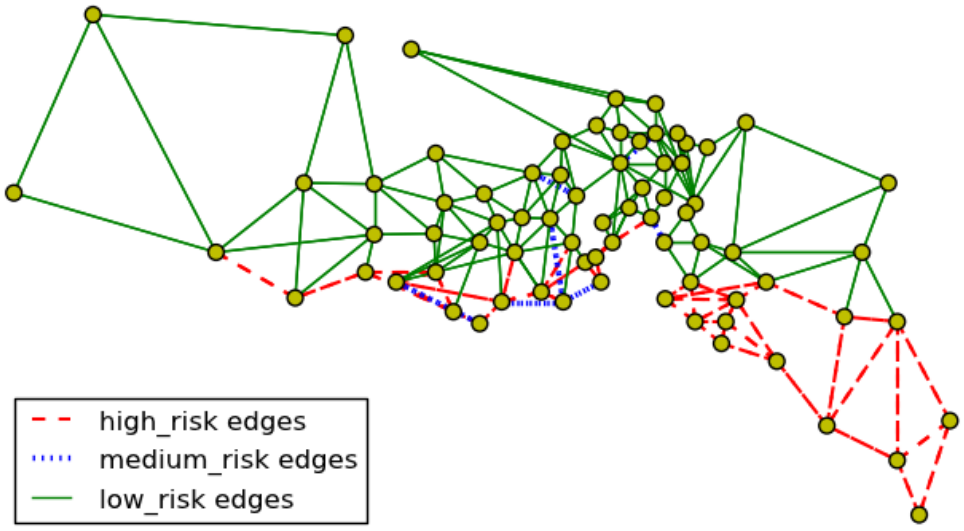


Figure 3 Categorization of Istanbul's roads.

solved instances. As we see when the number of vehicles gets higher the timing conflict cases gets more as we expected. All these instances were solved in less than 12 minutes. For the cases with one vehicle these run times were less than a minute but as the number of vehicles gets higher, it affects the run time accordingly.

7 Conclusion

We defined a new arc routing problem with the motivation of planning road clearance operations after a disaster. In this problem, we optimize the routes of K vehicles that traverse arcs and open a subset of blocked ones to reconnect the post disaster road network. We find which edges to unblock and walks of k vehicles, such that the longest walk is completed in minimum time. We call this problem *min max K-arc routing for connectivity problem(K-ARCP)*. We developed a heuristic approach that converts the solution of a relaxed MIP model to a feasible solution.

In spite of the difficulty of *K-ARCP*, we could derive the optimal solution in more than 60% of the tested cases by our approach. In fact for the case of one vehicle, our mathematical model solves *K-ARCP* exactly in all the instances. When the number of vehicles is small (at most 4), and number of nodes is large (≥ 100), the possibility of timing conflicts gets lower and results in better bounds for the optimal solution. On the other hand, when we put too many vehicles in a network with modest number of nodes, we may not use all the vehicles. In reality, there would be a very high cost associated with providing vehicles, since these vehicles in our problem are teams of machinery with required personnel and equipment. Therefore, fewer number of vehicles would be used for relatively smaller networks which will lower the possibility of time conflicts in the relaxed model's solution.

Exact methods to solve *K-ARCP* for small instances can be explored as future work. Other objectives that aim to connect the network partially with a given time limit could also be considered.

Table 5 Results of Istanbul's Network.

Number of Vehicles	Instance Number	Maximal Optimality Gap	Exact Optimal Solution	Optimal Value of *R-MIP* (in hours)
1	1	0 %	Yes	11.72
	2	0 %	Yes	6.99
	3	0 %	Yes	10.26
	4	0 %	Yes	12.15
	5	0 %	Yes	5.47
2	1	0%	Yes	5.83
	2	NS	–	–
	3	8 %	No	12.383
	4	10 %	No	17.05
	5	0 %	Yes	19.23
3	1	NS	–	–
	2	22 %	No	5.39
	3	NS	–	–
	4	25 %	No	4.62
	5	NS	–	-
4	1	NS	–	–
	2	NS	–	–
	3	NS	–	–
	4	NS	–	–
	5	NS	–	–

References

1 D. T. Aksu and L. Ozdamar. A mathematical model for post-disaster road restoration: Enabling accessibility and evacuation. *Transportation Research Part E: Logistics and Transportation Review*, 61(0):56–67, 2014.

2 A. N. Asaly. Logistics planning for restoration of network connectivity after a disaster. Master's thesis, Koc University, 2013.

3 A. N. Asaly and F. S. Salman. *Global Logistics, New Directions in Logistics Management*, chapter Arc Selection and Routing for Restoration of Network Connectivity after a Disaster. Taylor and Francis, 2014.

4 A. Assad, W. Pearn, and B. Golden. The capacitated Chinese postman problem: Lower bounds and solvable cases. *American Journal of Mathematical and Management Science*, 7:63–88, 1987.

5 E. Benavent, A. Corberán, I. Plana, and J. M. Sanchis. Min-max k-vehicles windy rural postman problem. *Networks*, 54(4):216–226, 2009.

6 A. Corberán, I. Plana, and J. M. Sanchis. A branch and cut algorithm for the windy general routing problem and special cases. *Networks*, 49(4):245–257, 2007.

7 J. Edmonds and E. L. Johnson. Matching, Euler tours and the Chinese postman problems. *Mathematical Programming*, 5:88–124, 1973.

8 H. A. Eiselt, M. Gendreau, and G. Laporte. Arc routing problems, Part II: The rural postman problem. *Operations Research*, 43(3):399–414, 1995.

9 C. S. Orloff. A fundamental problem in vehicle routing. *Networks*, 4(1):35–64, 1974.

10 W. L. Pearn. Solvable cases of the k-person Chinese postman problem. *Operations Research Letters*, 16:241–244, 1994.

11 Z. Win. *Contributions to Routing Problems*. PhD thesis, University of Augsburg, 1987.

12 S. Yan and Y.-L. Shih. Optimal scheduling of emergency roadway repair and subsequent relief distribution. *Computers and Operations Research*, 36:2049–2065, 2009.

A Review of Dynamic Bayesian Network Techniques with Applications in Healthcare Risk Modelling

Mohsen Mesgarpour[1], Thierry Chaussalet[2], and Salma Chahed[3]

1 **HSCMG, Faculty of Science and Technology, University of Westminster**
115 New Cavendish Street, London, W1W 6UW, UK
mohsen.mesgarpour@my.westminster.ac.uk

2 **HSCMG, Faculty of Science and Technology, University of Westminster**
115 New Cavendish Street, London, W1W 6UW, UK
chausst@westminster.ac.uk

3 **HSCMG, Faculty of Science and Technology, University of Westminster**
115 New Cavendish Street, London, W1W 6UW, UK
S.Chahed@westminster.ac.uk

Abstract

Coping with an ageing population is a major concern for healthcare organisations around the world. The average cost of hospital care is higher than social care for older and terminally ill patients. Moreover, the average cost of social care increases with the age of the patient. Therefore, it is important to make efficient and fair capacity planning which also incorporates patient centred outcomes. Predictive models can provide predictions which their accuracy can be understood and quantified. Predictive modelling can help patients and carers to get the appropriate support services, and allow clinical decision-makers to improve care quality and reduce the cost of inappropriate hospital and Accident and Emergency admissions. The aim of this study is to provide a review of modelling techniques and frameworks for predictive risk modelling of patients in hospital, based on routinely collected data such as the Hospital Episode Statistics database. A number of sub-problems can be considered such as Length-of-Stay and End-of-Life predictive modelling. The methodologies in the literature are mainly focused on addressing the problems using regression methods and Markov models, and the majority lack generalisability. In some cases, the robustness, accuracy and re-usability of predictive risk models have been shown to be improved using Machine Learning methods. Dynamic Bayesian Network techniques can represent complex correlations models and include small probabilities into the solution. The main focus of this study is to provide a review of major time-varying Dynamic Bayesian Network techniques with applications in healthcare predictive risk modelling.

1998 ACM Subject Classification G.3 Probability and Statistics

Keywords and phrases Healthcare Modelling, Dynamic Bayesian Network, Predictive Risk Modelling, Time-Varying, Hospital Administrative Data

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.89

1 Introduction

Costs of care are increasing at a rate that is unaffordable in the current economy. This is mainly due to the impact of ageing population, population growth, deprivation, increased expectations and cost of treatment and technology [45, 19]. The current system is unsustainable and unfair, and the current financial options to support people in meeting care costs are limited.

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 89–100

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In the fourth consecutive year, in 2013/14 local authorities have reduced social care budget [46]. Looking further ahead, it is projected that people aged over 85 to almost double by 2030, with an additional 600,000 of ageing population will be needing significant care [44]. While a quarter of people aged over 65 will need to spend very little on care over their life, half can expect cost of up to £20,000, and one in 10 can expect cost of £100,000 [44]. The fairness and quality of care can only be guaranteed if the cost saving changes are sustainable in the way care is provided. Because, there is some evidence that health and social care act as substitutes, pressure on one system can affect the other [27].

The main content is divided into seven sections. In the second section, the background is described briefly. The Predictive Modelling (PM) with applications to risk adjustment and predictive risk modelling is explained in the third section. The fourth section highlights the techniques in healthcare PM with emphasis on Dynamic Bayesian Network (DBN) modelling. Then, the main approaches in modelling time-varying DBN are highlighted in the fifth section. In the sixth section, the gaps and limitations of the current solutions are outlined. In the seventh section, the challenges are listed. Finally, in section eight, the future research areas are stated.

2 Background

Identification of high risk events has been a major concern for many healthcare providers and purchasers. According to Lewis et al. [43], there are three major sources of risks to the healthcare systems:

- Ageing population;
- Increase in number of people with long-term conditions;
- Rise in rate of emergency (or unplanned) admission to hospitals.

Firstly, a major concern in healthcare organisations throughout the world is about coping with an ageing population [10]. Survey results [12, 42] show that many people would prefer to die with an appropriate support at home rather than hospital, and yet the number of death in hospital can reach to 65% in the future if no appropriate policy would be in place.

Moreover, the average cost of hospital care is higher than social care for older and terminally ill patients. Also, the costs of care in the final phases before death are very high in hospital. According to the Nuffield Trust report [27], the average cost of social care increases with the age of the patient. However, the cost of social care stays cheaper than the hospital care for those whose age is below 85. Based on gender, the intersection point of hospital and social care costs for male patients are close to the age 90 and for the female patients are approximately at the age 80.

Sometimes a hospital admission can be avoided by residential setting substitution or social care. According to the analysis by Bardsley et al. [5] on a wide population in England, use of a social care may prevent the need for the hospital care. The End-of-Life researches help patients to get appropriate support services towards the end of life by better management of resources and patients. The ambition of National Health Service (NHS) is to increase the number of people who die in their usual place of residence to 60% . This baseline in 2007 was 38%, and with the End-of-Life practices that are in place it reached to 42% in 2012 [51, 52].

Secondly, ageing population and change of lifestyles are related to increase in number of people with long-term conditions and morbidity conditions (any condition that represent departure from a state of well-being). For instance, there was a significant rise in chronic kidney disease, diabetes and cancer between years 2006 and 2011 [18]. Also, it has been

predicted that the number of people with co-morbidity condition (dual diagnoses) to rise from 1.9 million in 2009 to 2.9 million in 2018 [47].

Thirdly, an estimated £11 billion per year is the cost of unplanned admissions to the NHS in England. There are two main causes of emergency admissions: early discharge and unpredictable accidents and emergency events. Based on a retrospective study by Clarke et al. [9], about half of the 30-day emergency readmissions were potentially preventable between 2004 and 2010. Discharging patients is a primary way of providing free beds in the healthcare systems, and administrators are in charge of evaluating the risk of emergency readmission in advance. Patient flow modelling such as Length-of-Stay PM enables managers to better understand the operational and clinical functions [2]. Length-of-Stay PM includes capturing the flow of patient from the admission to discharge. The flow is through a number of conceptual (virtual) phases that each patient goes through them. The PM of Length-of-Stay uses the time spent in phases in addition to the clinical data and the demographic data to identify events.

Therefore, it is important to make efficient and fair capacity planning. Predictive risk models that include the patient centred outcomes, as well as the aggregated levels, can help patients and carers to get the appropriate support services in clinical decision making. Also, PM of risks can improve care quality and reduce the cost of inappropriate hospital and Accident and Emergency (A&E) admissions.

The aim of this study is to provide a review of DBN techniques in healthcare predictive risk modelling, such as Length-of-Stay and End-of-Life PM.

3 Predictive Modelling

PM is often associated with Machine Learning (ML), pattern recognition and data mining. The practice of PM defines the process of development of models that their prediction accuracy can be understood and quantified [38]. Geisser [25] defined PM as "the process by which a model is created or chosen to try to best predict the probability of an outcome."

In the field of decision analysis, risk modelling [22] can answer questions such as:

- Is the risk high or low, and is the level acceptable?
- Which are the most critical causal factors?
- What are the differences between the risks of the alternative solutions?
- What can be achieved in terms of risk-reducing effect in comparison with the other options?

Generally, the identification of emergent risk can be categorised into modelling of three main aspects: stratification, clinical profiles and resource utilisation profiles. Moreover, in modelling of the events, the time dimension is usually modelled as time-to-event or as a risk score.

For instance, Gotz et al. [29] at the Watson Research Centre introduced an approach based on patient similarity, which extract a cohort of patients from an Electronic Health Record that is similar to specific patient, and provides interactive visualisation for complex decision making. It has been shown that this clustering, dynamic visualisation and expert refinement had been effective for near-term prognosis [21] and risk evaluation [11] of physiological data.

There are two major branches of risk modelling in healthcare: PM and Risk Adjustment. PM is frequently used for finding high-risk member Case Finding, such as finding patients with high risk of readmission and also to predict member cost and utilisation.

Risk Adjustment is a normalisation technique for comparison purposes, such as classifying patients by potential risk level for the purpose of the insurance provider reimbursement [43, 33]. These two methods are briefly summarised in the following subsections.

3.1 Predictive Risk Modelling

In healthcare systems various types of scoring systems (e.g. Glasgow Coma Scale for patients with brain injury) are used to support clinical and administrative decisions; however, statistical and stochastic models are needed to estimate the risks according to the changes in the care and environmental variables [60].

Physicians are interested in the evaluation and forecast of the adverse events that may provoke mortality or longer hospital stay for the patient, and assign a quantity to the patient's risk profile [14]. In terms of risk impact, healthcare risk analysis can be categorised into two categories: Operational Risks and Clinical Risks [36].

Data mining in healthcare can predict risks in healthcare, improve health status of high-risk patients and make overall savings. There are various predictive risk models in the literature, but each can forecast a small range of healthcare and social care outcomes. They differ in terms of the predicted time range, variables, data sources and the modelling approaches [43].

For instance, Billing et al. [6] developed a predictive risk model for re-admission using a multivariate logistic regression based on Length-of-Stay, diagnoses and demographic data derived from the Hospital Episode Statistics (HES) database. The algorithm produces risk scores for each admitted patient and a prediction of people in risk of readmission within 30 days. Also, 20 different risk bands were defined and associated to cost estimates for the business case analysis.

3.2 Risk Adjustment

Although the medical advances have contributed to improvement of life expectancy, it has little to do with the life expectancy and has much more to do with the quality of life. Risk Adjustment methods are either used for direct selection by health insurer for selection of good (profitable) risks from an insurer pool, or indirectly by designing insurance products. The models are often based on a linear utility function framework [48], and the objective is to minimise the outcome (risk) [15, 16]. The data mining techniques in Risk Adjustment modelling help to find the characteristics that have predictability power, which do not result into unfair risk assessment.

4 Techniques

Five major modelling methodologies presented in the previous studies on PM in healthcare are: simulation, formula-based methods, statistical, probabilistic and queuing. The focus of this study is on Dynamic Bayesian Network (DBN) techniques in PM, therefore initially regression modelling approach is briefly summarised. Afterwards, a brief introduction to the ML methods is provided. Then, the graphical networks with the main focus on DBN modelling are reviewed.

4.1 Regression

Regression modelling methods, such as logistic regression and mixed models, have been applied extensively in previous literature in social science and healthcare modelling [24, 39, 2].

An application of regression modelling is in the pathway modelling of the End-of-Life and frailty (i.e. the factors that arise from heterogeneity amongst patients) function modelling. A multinomial logit model was developed by Adeyemi et al. [1] for modelling patient's pathway. In the model, the patient frailties were regarded as mixed effect, and the random effects

distributions were modelled based on patient pathways. The model could identify the high probability pathways for survival and cost objective functions.

An evaluation of multiple logistic regression-based PMs was performed by Aylin et al. [4] for inpatient death using the HES database. Different number of factors, such as age, sex, admission parameters and diagnostics, were included, and ultimately the performance was compared by using Receiver-Operating Characteristic (ROC). The models were designed for very specific mortality risk index in healthcare, but it was demonstrated that the use of only an administrative database could effectively predict the risk.

4.2 Machine Learning

Since late 1980s, ML methods have been used in extending the statistical analysis for making inferences from data. There are a lot to be done in the area of automated methods for learning and forecasting in healthcare. In this research, the major area of contribution is in developing techniques for doing transfer learning. The transfer learning refers to the methods that harness and adapt models to a specific new predictive task at hand. Transfer learning methodologies can help to use forecasting and PM to provide a systematic methodology of analysis for similar cases with smaller number of visible parameters. This may also be extended to perform active learning for use in complex real-world settings [34, 30].

Based on the knowledge of interest, Bayesian Networks (BNs), Neural Networks, Decision Tree and Kernel methods, such as Support Vector Machines, and Gaussian Processes are often used in healthcare data mining. Other approaches in ML can be found in the work of Bishop and Nasrabadi [7]. In the following, BNs which is a subset of graphical networks is discussed.

4.2.1 Bayesian Networks

There are two main approaches in incorporating stochastic models to the statistical modelling: discriminative (conditional distribution model) and generative (joint probability model). The discriminative modelling does not make any assumption about the prior distribution and only includes the conditional probability. Therefore, discriminative modelling is also known as the frequentist approach, and a linear classifier and logistic regression are examples of it. On the other hand, Bayes modelling methods are known as generative and they include the prior (marginal) distribution of the evidence data to the discriminative model [49].

Graphical networks are commonly used in probabilistic and statistical modelling. BNs [54] are standard approaches for modelling structured domains by allowing explicit representation of dependencies. BNs use probabilistic inference methods to present parameters with high dimensional distributions. These parameters are presented as nodes in graphical networks, where links between nodes represent direct correlation. To be able to update and infer large networks' probabilities in real-time, inference approximation methods, such as Expectation Maximisation and Particle Filter, have been introduced to estimate the joint and marginal probabilities of the nodes.

In BN modelling, template-based representations are used to produce a single compact model that can represent properties of system dynamics and to produce distribution over the different trajectories (DBN modelling) or to produce a distribution over different worlds (e.g. Genetics networks). To be able to reason about non-static situations, DBNs [17] are used to represent nodes with system-states. The system-states are either considered as stationary time-slices (homogeneous or invariant) like Markov Models or as the state observation model like Hidden Markov Models (HMMs). In state observation models, the states are variant and evolve on their own separately from the observations [37].

There are always two major challenges in using BN modelling: design and training. The accuracy and the efficiency of BN models depend on four main design choices: the framework of the causal tree, the framework of the system-states, inference approximation algorithm and finally the assignment and update method of prior probabilities.

It is not feasible computationally to design a graphical network that is too large, dynamic, inhomogeneous, noisy and incomplete. Therefore, the modelling assumptions are often relaxed, and specialised approximation and heuristic methods are widely used. In the next section, five main Bayesian Network (BN) techniques in PM of dynamic systems are outlined.

5 Time-Varying Dynamic Bayesian Networks Approaches

In a PM problem, such as Length-of-Stay or End-of-Life, the features are mainly inhomogeneous, because the processes in the models are either non-stationary (e.g. length of illness or treatment stages) or the events are sparse (e.g. morbidity conditions or patient states). The unobservable and inhomogeneous properties of models cause the momentum of the system dynamic to change across temporal access.

Generally, it is not statistically tractable to consider all of the variances for every timepoint, because of the complexity in the inference and the lack of enough training evidence. Also, it is not possible to segment the time, since the model characteristics are unknown for each segment. There are five main approaches to model a time-varying Dynamic Bayesian Network (DBN). The approaches are highlighted in the following and the summary of the studies is presented in Table 1.

Firstly, a basic indirect approach is to transform time in order to make the process homogeneous. A naive approach is to use a time interpolation technique. Instead of using direct time transformation, a method like Kalman filter can be used, which is a Linear Dynamical System technique and is based on an autoregressive function to estimate a value at a timepoint [62, 13]. Xu et al. [63] proposed a state space model based on Kalman filter to estimate mean and variance for equally and unequally spaced longitudinal count data with serial correlation. The model applied to Epileptic Seizure and Primary Care Visits Data, and with high number of observations, the model produced comparable results to those by numerical approach.

Moreover, another indirect approach is to re-weight the likelihoods at each time-point using a particle based approach, such as feed-forward and sparse Kalman filtering. Since DBN is a generative model, it often works better for sparse models, because of its assumption about the underlying probabilities. However, it needs to be applied with an extreme care, since inappropriate sampling technique can rapidly slide the weights to zero or make the model assumptions and prior probabilities incorrect [37].

Furthermore, another approach is network rewiring which is also known as time-evolving graphical models [57]. It is a feasible option for large-scale time-varying networks. Time-evolving graphical models have been recently used in designing large networks in biological and social studies [3, 64, 31], with the objective to find unobserved network topologies or to rewire network under different conditions (e.g. edge-stability, and transitivity). For instance, recently a new modelling approach known as temporal Exponential Random Graph Model has been proposed for modelling networks evolving over discrete time-steps with Monte Carlo Markov Chain based or convex optimisation algorithms for posterior inference [3, 31, 32].

Another approach is conditional BN modelling, which is also known as multilevel or hierarchical BN model and is popular in the literature. In BN modelling a time-varying framework on top of a Markov Chain (MC) technique can be used to model multilevel time

properties. A Cox phase-type model is used for modelling durations on top of a Hidden Semi-Markov Model by [20] for human activity recognition. In this research, an extension added to the Coxian Hidden Semi-Markov Model, which incorporates both duration and hierarchical modelling. Also, a DBN framework is proposed by Lappenschaar et al. [41] for modelling non-stationary events in multi-morbidity modelling. This PM of the interactions between heart failure and diabetes mellitus could closely resemble the PM techniques which use multilevel Linear Mixed Model. Moreover, a framework is designed by Lappenschaar et al. [40] for formulating an Linear Mixed Model into a BN using a Logistic Regression function.

Finally, the Linear Dynamical Systems are useful temporal models, which represent one or more real-valued variables that evolve linearly over time, with some Gaussian noise [37]. There are two categories of the Linear Dynamical Systems methods in modelling of time-varying DBN: Switching Linear Dynamic System and Time Varying Autoregression.

Switching Linear Dynamic Systems have been studied extensively for piecewise modelling of linear systems [61]. Based on the Switching Linear Dynamic Systems modelling, time-varying observations [28] and time-varying duration [8, 53] can be formulated using a latent MC. But, the MC method which is a piecewise stationary, does not have a very general application in learning and inference, and a time-varying linear regression can be used instead. For instance, a time-varying DBN has been introduced by Song et al. [59], which aggregates observations of adjacent time points by a kernel re-weighting function.

Time Varying Autoregression models are another type of the Linear Dynamic Systems [61] models, which focus on non-stationary models with fixed structure. Time Varying Autoregression models have been applied on a wide range of research applications, such as PM of equity market [35], inferring time-varying data from gene expression [56, 55] and modelling non-Gaussian autoregression [26].

6 Gaps and Limitations

Presented solutions in the literature mainly lack robustness and re-usability in different environments and for different care plans. These approaches mainly use descriptive, regression or Markov methods. The main shortcomings of these models can be summarised in the following:

- Not being fit for the modelling of complex correlations;
- Not accounting for small probabilities in an appropriate way;
- Not updating the beliefs (prior probabilities) based on the environmental variables and change of policies;
- Not using an automated parallel workflow system to compare different models and settings [50, 22].

Moreover, after the breakdown of financial markets at 2008, Rodriguez [58] wrote, "PM, the process by which a model is created or chosen to try to best predict the probability of an outcome has lost credibility as a forecasting tool". This is due to either the modeller's expertise or knowledge, or the lack of resources. The main common reasons that make a PM model to fail are as below:

- Inadequate pre-processing of the data;
- Inadequate model validation;
- Unjustified extrapolation;
- Over-fitting the model to available data [38].

Table 1 The summary of the studies in modelling time-varying processes.

Approach	Method(s)	Study	Domain	Findings/Outcomes
Time transformation	Auto-regression and Kalman filter	Xu et al. [63]	Healthcare events	The likelihood estimation approach performs better than the numerical integration approach
Time Evolving Graphical Network	Prevailing networks	Robinson et al. [57]	Generic non-stationary data	Demonstrating the feasibility
	Scalable inference for time-evolving networks	Ahmed et al. [3]	Biological systems to social science	Having asymptotically value-consistent under fixed model dimension
Conditional BN Modelling	A multilevel BN	Lappenschaar et al. [40]	Multimorbidity condition prediction	Providing more insight into interaction of multiple diseases
	A multilevel BN	Lappenschaar et al. [41]	The course of a medical condition	An informative clinical decision making tool
	Semi-Markov model, Coxian and HMM	Duong et al. [20]	Recognition of human activities of daily living	Having high and comparable accuracy
Switching Linear Dynamic Systems	Using hidden variables for network changes	Wang et al. [61]	Camera tracking	Being successful for both simulated non-stationary data and video sequences
Time Varying Autoregression	Kalman filter	Johnson et al. [35]	Equity market prediction	Performing as well as the Capital Asset Pricing Model benchmark, despite of using non-traditional pricing measures

The PM of Operational Risks in healthcare modelling, such as Length-of-Stay and End-of-Life, varies across systems and often are not robust. There is a few examples in the literature that uses ML techniques including BNs to address Length-of-Stay and End-of-Life problems.

7 Challenges

There are a number of challenges in applying ML techniques in healthcare modelling. Firstly, there are uncertainties in patient flows and resource demands as well as complex interdependency relations in the healthcare system.

Secondly, there are parameters, associations and timeframes that are missing from the primary and the secondary database due to security and confidentiality concerns. Thirdly, understanding the parameters, dependencies and independences are very crucial. Fourthly, there are various challenges that lie ahead of the robustness and generalisation of the solution, such as supporting adequate training data and selection of performance measures [23, 4].

Finally, it is very common in PM to build, evaluate and compare different models with varying features, algorithms, parameters and cohorts. Therefore, use of an optimised and automated parallel workflow system, like the framework developed by Ng et al. [50] is desirable to facilitate large-scale modelling.

8 Future Research Directions

Available solutions and proposed methodologies in the literature are mainly focused on addressing healthcare problems, such as PM of Length-of-Stay and End-of-Life, using regression methods and Markov models on very small number of case studies. The robustness, accuracy and re-usability of the models can be improved using ML methods.

ML methods, such as BNs, can represent complex correlation models and include small probabilities into the solution. BN methodologies allow composing risk utility functions and the causal and dependency relationship between variables, events, risks and outcomes. This enables BN models to appropriately estimate small probabilities that are associated to healthcare system as well as the statistical evidence [22]. There has been a little research done on using BNs in healthcare modelling, and the time-varying DBN techniques have great potentials in PM of non-stationary and hierarchical events and risks in healthcare problems.

References

1 Shola Adeyemi and Thierry J Chaussalet. Models for extracting information on patient pathways. In *Intelligent Patient Management*, pages 171–182. Springer, 2009.

2 Shola Adeyemi, Eren Demir, and Thierry Chaussalet. Towards an evidence-based decision making healthcare system management: Modelling patient pathways to improve clinical outcomes. *Decision Support Systems*, 55, 1:117–125, 2013.

3 Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

4 Paul Aylin, Alex Bottle, and Azeem Majeed. Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *BMJ: British Medical Journal*, 334(7602):1044, 2007.

5 Martin Bardsley, Theo Georghiou, Ludovic Chassin, Geraint Lewis, Adam Steventon, and Jennifer Dixon. Overlap of hospital use and social care in older people in england. *Journal of health services research and policy*, 17(3):133–139, 2012.

6 John Billings, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, and Martin Bardsley. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30). *BMJ Open: e001667*, 2(4), 2012.

7 Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

8 A. Blake, B. North, and M. Isard. Learning multi-class dynamics. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 389–395, Cambridge, MA, USA, 1999. MIT Press.

9 Ian Blunt, Martin Bardsley, Amy Grove, and Aileen Clarke. Classifying emergency 30-day readmissions in england using routine hospital data 2004-2010: what is the scope for reduction? *Journal of Epidemiology and Community Health*, 66(Suppl 1):A45–A45, 2012.

10 Michael Caley and Khesh Sidhu. Estimating the future healthcare costs of an aging population in the uk: expansion of morbidity and the need for preventative care. *Journal of Public Health*, 33(1):117–122, 2011.

11 S Chattopadhyay, P Ray, HS Chen, MB Lee, and HC Chiang. Suicidal risk evaluation using a similarity-based classifier. In *Advanced Data Mining and Applications*, pages 51–61. Springer, 2008.

12 Xavier Chitnis, T Georghiou, A Steventon, and M Bardsley. The impact of the marie curie nursing service on place of death andhospital use at the end of

life. https://www.mariecurie.org.uk/Documents/HEALTHCARE-PROFESSIONALS/Our%20impact/Marie%20Curie_Full%20Report_Final_Web.pdf, 2012.

13 Richard J Cook and Jerald F Lawless. Statistical issues in modeling chronic disease in cohort studies. *Statistics in Biosciences*, 6:127–161, 2013.

14 Chiara Cornalba. Clinical and operational risk: A bayesian approach. *Methodology and Computing in Applied Probability*, 11(1):47–63, 2009.

15 Anthony J. Culyer and Joseph P. Newhouse. *Handbook of Health Economics: Vol. 1A*. Elsevier, 2000.

16 Anthony J. Culyer, Mark V. Pauly, Joseph P. Newhouse, Thomas G. McGuire, and Pedro P. Barros. *Handbook of Health Economics: Vol. 2*. Elsevier, 2012.

17 Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150, 1989.

18 DoH. Long term conditions compendium of information (third edition). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216528/dh_134486.pdf, Apr. 2012.

19 DoH. Business case: for the health and care modernisation transition programme. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/224140/TP_business_case_docx.pdf, 2013.

20 Thi Duong, Dinh Phung, Hung Bui, and Svetha Venkatesh. Efficient duration and hierarchical modeling for human activity recognition. *Artificial Intelligence*, 173(7):830–856, 2009.

21 Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics. In *AMIA Annual Symposium Proceedings*, volume 2010, page 192. American Medical Informatics Association, 2010.

22 Norman E Fenton and Martin D Neil. *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, 2012.

23 Jonathan Agner Forsberg, Rikard Wedin, Bjarne H Bauer, Henrik CF andHansen, Minna Laitinen, Clement S Trovik, Johnny Keller, Patrick J Boland, and John H Healey. External validation of the bayesian estimated tools for survival (bets) models in patients with surgically treated skeletal metastases. *BMC cancer*, 12(1):493, 2012.

24 G David Garson. *Hierarchical linear modeling: Guide and applications*. Sage, 2012.

25 Seymour Geisser. *Predictive interference: an introduction*, volume 55. CRC Press, 1993.

26 Deniz Gencaga, Ercan E. Kuruoglu, and Aysin Ertuzun. Modeling non-gaussian time-varying vector autoregressive processes by particle filtering. *Multidimensional Systems and Signal Processing*, 21(1):73–85, 2010.

27 Theo. Georghiou, Sian Davies, Alisha Davies, and Martin Bradley. Understanding patterns of health and social care at the end of life. http://www.nuffieldtrust.org.uk/sites/files/nuffield/121016_understanding_patterns_of_health_and_social_care_full_report_final.pdf, Oct. 2012.

28 Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000.

29 David Gotz, Jimeng Sun, Nan Cao, and Shahram Ebadollahi. Visual cluster analysis in support of clinical decision intelligence. In *AMIA Annual Symposium Proceedings*, volume 2011, pages 481–490. American Medical Informatics Association, 2011.

30 Susan Graham, Deborah Estrin, Eric Horvitz, Isaac Kohane, Elizabeth Mynatt, and Ida Sim. Information technology research challenges for healthcare: From discovery to delivery. *ACM SIGHIT Record*, 1(1):4–9, 2011.

31 Fan Guo, Steve Hanneke, Wenjie Fu, and Eric P Xing. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328. ACM, 2007.

32 Steve Hanneke, Wenjie Fu, Eric P Xing, et al. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

33 Dawn E Holmes and Lakhmi C Jain. *Data Mining: Foundations and Intelligent Paradigms: Volume 3: Medical, Health, Social, Biological and Other Applications*, volume 3. Springer, 2012.

34 Eric Horvitz. From data to predictions and decisions: Enabling evidence-based healthcare. `http://research.microsoft.com/pubs/141911/Evidence_based_healthcare_essay.pdf`, 2010.

35 Lorne D Johnson and Georgios Sakoulis. Maximizing equity market sector predictability in a bayesian time-varying parameter model. *Computational Statistics and Data Analysis*, 52(6):3083–3106, 2008.

36 Linda T Kohn, Janet M Corrigan, Molla S Donaldson, et al. *To err is human: building a safer health system*, volume 627. National Academies Press, 2000.

37 Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

38 Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

39 Elena Kulinskaya, Diana Kornbrot, and Haiyan Gao. Length of stay as a performance indicator: robust statistical methodology. *IMA Journal of Management Mathematics*, 16(4):369–381, 2005.

40 Martijn Lappenschaar, Arjen Hommersom, Peter JF Lucas, Joep Lagro, and Stefan Visscher. Multilevel bayesian networks for the analysis of hierarchical health care data. *Artificial intelligence in medicine*, 57, 3:171–183, 2013.

41 Martijn Lappenschaar, Arjen Hommersom, Peter JF Lucas, Joep Lagro, Stefan Visscher, Joke C Korevaar, and François G Schellevis. Multilevel temporal bayesian networks can model longitudinal change in multimorbidity. *Journal of clinical epidemiology*, 66(12):1405–1416, 2013.

42 Charles Leadbeater and Jake Garber. Dying for change. `http://www.demos.co.uk/files/Dying_for_change_-_web_-_final_1_.pdf`, 2010.

43 Geraint Lewis, Natasha Curry, and Martin Bardsley. Choosing a predictive risk model: a guide for commissioners in england. `https://www.primis.nottingham.ac.uk/attachments/article/643/choosing_predictive_risk_model_guide_for_commissioners_nov11.pdf`, 2011.

44 Funding of Care and Support. Fairer care funding - the report of the commission on funding of care and support. `http://www.ilis.co.uk/uploaded_files/dilnott_report_the_future_of_funding_social_care_july_2011.pdf`, 2011.

45 NHS England. Nhs england publishes ccg funding allocations for next two years following adoption of new formula. `http://www.england.nhs.uk/2013/12/18/ccg-fund-allocs/`, 2013.

46 The King's Fund. Briefing: The care bill - second reading in the house of commons. `http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/briefing-care-bill-house-of-commons-second-reading-kingsfund-dec13.pdf`, 2013.

47 The King's Fund. The health and social care system in 2025 - a view of the future. `http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/The%20Health%20and%20Social%20Care%20System%20in%202025%20-%20supplementary%20note%20from%20The%20King's%20Fund.pdf`, 2013.

48 Joseph P Newhouse. Reimbursing health plans and health providers: efficiency in production versus selection. *Journal of economic literature*, 34(3):1236–1263, 1996.

49 Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

50 Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. Paramo:a parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of Biomedical Informatics*, 2013.

51 NHS. Whole system partnership - end of life care engagement event. Presentation, Jul. 2012.

52 NHS. Improving end of life care through early recognition of need. `http://www.thewholesystem.co.uk/docs/eolc_predictive_modelling_report.pdf`, 2013.

53 Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. In *NIPS*, pages 981–987. Citeseer, 2000.

54 Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

55 Bruno-Edouard Perrin, Liva Ralaivola, Aurelien Mazurie, Samuele Bottani, Jacques Mallet, and Florence d'Alche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl 2):ii138–ii148, 2003.

56 Arvind Rao, Alfred O Hero III, James Douglas Engel, et al. Inferring time-varying network topologies from gene expression data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:7–7, 2007.

57 Joshua W. Robinson and Alexander J. Hartemink. Learning non-stationary dynamic bayesian networks. *J. Mach. Learn. Res.*, 11:3647–3680, December 2010.

58 Manuel A. Rodriguez. The failure of predictive modeling and why we follow the herd. `http://www.cfclaw.com/files/r_201145133147.pdf`, 2011.

59 Le Song, Mladen Kolar, and Eric P Xing. Time-varying dynamic bayesian networks. In *NIPS*, pages 1732–1740, 2009.

60 Roger Stedman. Scoring systems and outcomes. In Fang Gao Smith, editor, *Core Topics in Critical Care Medicine*, pages 27–33. Cambridge University Press, 2010.

61 Zhaowen Wang, Ercan E Kuruoglu, Xiaokang Yang, Yi Xu, and Thomas S Huang. Time varying dynamic bayesian network for nonstationary events modeling and online inference. *Signal Processing, IEEE Transactions on*, 59(4):1553–1568, 2011.

62 Zhu Wang, Wayne A Woodward, and Henry L Gray. The application of the kalman filter to nonstationary time series through time deformation. *Journal of Time Series Analysis*, 30(5):559–574, 2009.

63 Stanley Xu, Richard H Jones, and Gary K Grunwald. Analysis of longitudinal count data with serial correlation. *Biometrical journal*, 49(3):416–428, 2007.

64 Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *arXiv preprint arXiv:0802.2758*, 2008.

Demand models for the static retail price optimization problem – A Revenue Management perspective

Timo P. Kunz and Sven F. Crone

Department of Management Science, Lancaster University
Lancaster LA1 4YX, United Kingdom
{t.p.kunz, s.crone}@lancaster.ac.uk

Abstract

Revenue Management (RM) has been successfully applied to many industries and to various problem settings. While this is well reflected in research, RM literature is almost entirely focused on the dynamic pricing problem where a perishable product is priced over a finite selling horizon. In retail however, the static case, in which products are continuously replenished and therefore virtually imperishable is equally relevant and features a unique set of industry-specific problem properties. Different aspects of this problem have been discussed in isolation in various fields. The relevant contributions remain therefore scattered throughout Operations Research, Econometrics, and foremost Marketing and Retailing while a holistic discussion is virtually non-existent. We argue that RM with its interdisciplinary, practical, and systemic approach would provide the ideal framework to connect relevant research across fields and to narrow the gap between theory and practice. We present a review of the static retail pricing problem from an RM perspective in which we focus on the demand model as the core of the retail RM system and highlight its links to the data and the optimization model. We then define five criteria that we consider critical for the applicability of the demand model in the retail RM context. We discuss the relevant models in the light of these criteria and review literature that has connected different aspects of the problem. We identify several avenues for future research to illustrate the vast potential of discussing the static retail pricing problem in the RM context.

1998 ACM Subject Classification G.1.10 Applications

Keywords and phrases Revenue Management, Pricing, Retail, Demand Modeling

Digital Object Identifier 10.4230/OASIcs.SCOR.2014.101

1 Introduction

The practical value of Revenue Management (RM) and its success throughout many industries is well documented. RM can also be lauded for its achievements as an academic discipline as it has been tremendously successful in bridging the gap between theory and practice and connecting industry and academia. Even more importantly, it has established itself as a discipline in its own right that encourages a broader view on problems usually only studied in isolation, bringing together aspects from various fields ranging from forecasting, econometrics and mathematical programming, to computing, strategic management, operations management, and behavioural marketing.

While RM takes different perspectives (e. g. quantity vs. price based or industry specific), the commonality typically is the scope of dynamically pricing perishable products over a finite selling horizon. However, in most retail formats, the static problem of pricing products that are continuously replenished and virtually non-perishable is fundamental and constitutes

4th Student Conference on Operational Research (SCOR'14).
Editors: Pedro Crespo Del Granado, Martim Joyce-Moniz, and Stefan Ravizza; pp. 101–125

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

a problem in its own right. This has been insufficiently acknowledged in RM literature where the dynamic problem is ubiquitous and the static case, as it presents itself in the retail scenario, is only sparsely recognized: While most RM textbooks do not address the problem at all (e. g. [98], [63]), Talluri et al. [109] very briefly discuss the special characteristics of the grocery pricing case and mention the determination of "baseline prices". In their survey on emerging trends in retail pricing practice, Levy et al. [72] acknowledge "two disparate pricing problems: Fashion and staple merchandise" and discuss the static pricing problem for the latter.

We argue that studying static retail pricing under an interdisciplinary proposition offers interesting opportunities in research and practice alike, as well as for the convergence of the two, and promises to be equally rewarding as it has been for the dynamic problem. While there is a considerable body of surveys for the dynamic scenario where many also address the dynamic retail case ([13], [27], [45], [39], [122], [83]), a review that frames the static retail price optimization problem in this way does not exist.

We want to advance the discussion by filling this gap and discussing the demand models used for the static retail price optimization problem in an RM context. In this spirit, we take an interdisciplinary view, highlighting the interplay of the optimization system components – in particular the data, demand, and optimization model – and the questions emerging from the interactions of these components. We define five criteria from a practical point of view that we consider critical for the relevance of the demand model for real life applications and discuss existing demand models under this premise. We argue that it is beneficial to include the static problem in the RM discourse and develop directions for future research.

We proceed as follows: In Section 2, we give a broad overview of where the static pricing problem has received attention and discuss contributions that take a systemic view in an RM sense. In Section 3, we highlight the characteristics of the problem that originate from the specific properties of the retail environment and define five criteria that we consider essential for applicability. Subsequently, in Section 4, we review demand models which we organise in absolute and relative models. In Section 5, we discuss these models with respect to the criteria defined earlier. We conclude by developing suggestions for future research in Section 6.

2 Literature on static retail pricing

2.1 The scope of Revenue Management literature

RM as a field is designed to be interdisciplinary. This is illustrated by the standard text books of the field that cover diverse aspects of the RM process, such as demand model estimation, optimization, forecasting, behavioural aspects, or implementation (e. g. [109], [98], [63]). Talluri et al. [109, p.5] argue in their text that the scientific advances in Economics, Statistics, Operations Research (OR) and Information Technology are driving the RM approach. Interdisciplinarity is also a cornerstone in the declared scope of the journals dedicated to the field (e. g. Journal of Revenue and Pricing Management, International Journal of Revenue Management) that also emphasize applicability and stress their ambitions to bring practice and academia together. Dedicated RM articles covering questions around the 'traditional', dynamic RM problem also exist side by side in the literature that discusses the theoretical groundwork for these contributions, most notably in Operations Research and Management Science where RM holds a strong footprint.

The static retail pricing problem is virtually not included in this discussion. Questions around retail pricing and the optimal price are discussed within Marketing and Retailing

instead. While the relevant publications (e. g. Journal of Retailing, Marketing Science) regularly feature contributions that are useful in the context, RM as a discipline is here practically non existent. Relevant articles on pricing topics, a very popular research stream, are usually discussed under terms such as 'product line pricing', 'pricing and promotion', or 'optimal pricing'. The difference in naming is of course not an issue, however RM has gone a long way in facilitating an interdisciplinary and practical discussion as described above which currently does not exist for the static problem.

The above can be further illustrated by the challenges and suggestions for future research presented in dedicated reviews of the different fields. In RM literature, the research proposed seems to span a wide variety of themes that include aspects not directly associated with typical OR interests and often driven by very practical concerns: reoccurring themes are processing and operationalisation of data, often with links to forecasting and predictive accuracy, concerns of decision support systems, including implementation, and the (probably more typical) overarching question of how to increase the accuracy, efficiency and general benefit of the system and its components (e. g. [27], [73], [39]). While the avenues for future research suggested in similar contributions in the Marketing and Retailing literature point to equally important questions, a systemic view is rarely considered. The focus seems more rooted in the respective discipline and often centered around issues of customer behaviour, such as the acceptance of and the reaction to (optimal) prices and promotions, price implementation from an organisational perspective, and effectiveness from a behavioural point of view (e. g. [72], [47]).

2.2 Retail pricing literature in the Revenue Management context

Research that examines the static pricing problem in this way has been undertaken. However, contributions are scattered throughout disciplines and hence do not form a coherent discussion. In the following we want to give a brief overview where the relevant studies can be found and link to questions and articles for the three connecting fields demand modeling, optimization, and decision support systems. Rather than providing an exhaustive review, we want to highlight a few selected studies that can provide a starting point for a more comprehensive review.

The largest body of literature relevant for the problem can be found under the term 'product line pricing'. It is a rather wide field, mainly motivated by demand interdependencies, that includes the small scale manufacturer's as well as the large scale retail problem (reviews in [49], [84], [99]). However, in Marketing and the affiliated Retail literature the discussion of product line pricing has been accelerated by the adoption of the Category Management idea. The primary focus of these studies is usually price elasticities and their purpose is often purely descriptive (e. g. [24], [18], [102], [14], [55]). In a time where the only normative guidance offered was in the form of pricing heuristics, Urban [113] presented one of the first works to use econometric instruments to explicitly address product line pricing decisions with notable contributors to follow in the decades to come (e. g. [100], [26], [127], [119], [85], [87]). In a parallel research stream, similar models have been been developed and used for the promotion problem (e. g. [111], [15]).

First, we want to highlight research at the intersection of data and demand modeling. As this is obviously not a specialty of Marketing, few descriptive and almost no normative contributions focusing on such questions can be found. The influence of data conditions (e. g. seasonality, stationarity, intermittency) on the problem at hand, a topic foremost treated in the forecasting literature, has not been evaluated yet. Some papers have addressed questions of data processing (e. g. data pruning, data aggregation, data cleansing) in a

descriptive way (e. g. [4], [19], [126]). Questions around data types and data availability have been discussed similarly (e. g. [40], [54]). Directly intertwined with the above is the development of corresponding estimation methods, a topic usually primarily treated in the corresponding Econometrics literature. Most visible in our context have been contributions based on Bayesian methods (e. g. [58], [101]).

Second, unlike for the dynamic case, there is no natural link to OR literature where the intricacies of the resulting optimization problem are discussed. The literature dedicated to special questions arising from that context appears especially sparse. The only contribution explicitly dedicated to the problem that has come to our attention is Subramanian et al. [108] who present a linearization for an MNL based retail price optimization model. In some instances, the static pricing problem can also be found treated jointly with questions from inventory management or assortment planning (review e. g. [65]). Here, the literature covering aspects of optimization seems significantly more developed.

Third, some studies that focus on operationalisation of the models from a Decision Support System view or in the form of practice reports exist. Naturally, commercial providers of optimization systems do not bring their proprietary models into the public domain. Exceptions are some applications with commercial implementations that have been driven by academia and are hence published. Most notable are a price optimization system for a DIY retailer ([94], [95], [93]), and a system for treating the very unique problem of an automotive aftermarket retailer [78]. Some systems are documented that cover promotion planning but are insightful for the static problem as well (e. g. [12], [38], [105]). Further, some contributions highlight aspects of the architecture of such a system (e. g. [41]) or address questions around their implementation (e. g. [86]).

3 Framing static retail pricing in the Revenue Management context

Many models rest on assumptions or simplifications that can not be upheld when considering the reality of the retail environment. While studying these models undoubtedly holds academic merit, in this review we want to focus on, and accentuate applicability. It is therefore essential to consider the defining characteristics of the static retail problem which are decisive for the practical suitability of the model. In agreement with Kopalle [66] who names factors to consider when optimizing retail merchandising decisions and Levy et al. [72] who determine aspects to consider for determining optimal prices in retail, we focus entirely on the interaction between data, demand, and optimization model in the RM system and define five requirements that a demand model needs to satisfy in order to be applicable in a real life retail environment:

1. Inclusion of cross price effects
2. Reliance on operational (store-level) data only
3. Suitability for industry size problems
4. Potential to accommodate typical retail data conditions
5. Sensible solutions when used for optimization

(1) Inclusion of cross price effects: The demand model can be considered to be the core of the price optimization system [118]. Within the system, it acts as the middle piece between data and optimization model and is also the centre of our analysis. As the first requirement, we want to discuss a characteristic that is demand model intrinsic: The interdependency between products is a defining property of the retail assortment and a key-differentiator to the equivalent manufacturer problem. The consideration of cross price effects in the

demand model formulation is therefore fundamental. If this is relaxed, which effectively implies ignoring the assortment context and optimizing the price of each product in isolation, the problem is greatly reduced but clearly leads to inferior results (e. g. [100], [46]). Even though we acknowledge that systems that only rely on an abstract or indirect consideration of cross price effects have also proven to be useful (e. g. [95], [93]), we formulate as our first requirement that the demand model needs to include cross price effects.

(2) Reliance on operational (store-level) data only: We want to highlight the sources of retail data. Two forms of data are commonly discussed: Store-level data, that usually contain sales and marketing mix information per item and store, most often aggregated to a weekly level, and panel data, which are on a household level and therefore tend to include additional information such as demographics. While both have individual advantages and disadvantages and often have complementary uses, there is no consensus in research whether either will actually deliver superior results or affect model accuracy ([110], [1], [51], [5]). However, for applied purposes, store level data can be considered far more relevant as they are readily available as a byproduct of the payment process and therefore inexpensive to acquire. Data that contain additional details such as in-store product location, store layout, or manufacturer marketing efforts can normally not be reliably extracted from operative systems and, even though potentially usable in experimental settings, are largely unsuited for a productive RM system with the capability to address managerial issues on a realistic scale. Further, the cost and effort to collect panel data make them prohibitive for operational use. Even though a lot of retailers have customer card schemes that allow them to capture similar, panel like data, due to the self selective nature of these schemes that usually only appeal to a specific group of customers, the data generated are not representative and only partially useful for our purposes. Therefore, as a second requirement, it needs be possible to estimate the demand model given store-level scanner data that can be captured from operational processes.

(3) Suitability for industry size problems: We want to discuss the challenge of estimating the demand model given the empirical dimensions of the retail problem described above. The data reality of retail is usually defined by its high volume as can be seen by the dimensions of data sets such as the Dominick's database [61] or the IRI marketing data set [22]. It involves thousands of products in an assortment in which even the smallest category has more than a hundred individual products, thousands of transactions and often a great number of stores. Levy et al. [72] even describe the sheer size of the problem as "daunting".

Therefore, as a third requirement, it needs to be possible to estimate the model on a realistic, industry size scale without reducing or oversimplifying the data, in a way that the capability of the system to determine prices on the level of the individual product remains uncompromised.

(4) Potential to accommodate typical retail data conditions: We take a closer look at conditions and properties of such data in the form that they are actually found in real data scenarios as it will further complicate the estimation problem described above. Additional challenges for the estimation of the demand model originate from underlying data conditions: Examples are dynamic variation of the choice set over time due to changing assortment through product introduction and discontinuation, strong seasonality, demand intermittency for slow moving items, or seemingly intermittent demand due to stock out situations. It is unreasonable to expect to effortlessly accommodate every possible data condition in a

demand model, yet it is essential that the formulation of the model is flexible enough that it can be adjusted accordingly or that it allows for estimation techniques fit for the data conditions encountered. We purposely choose a broad formulation for our fourth requirement and state that the demand model needs to allow estimation under diverse data conditions.

(5) Sensible solutions when used for optimization: Finally, we want to take a closer look at the interaction between the demand model and the optimization model. As mentioned above, the presence of cross price effects is a defining characteristic of the retail assortment. This has implications for the determination of optimal prices: commonly, products in a category are substitutes rather than complements and their cross price elasticities are therefore positive. Due to this property, popular demand model forms such as linear or log-linear formulations, or choice based market share models imply that maximum profit is obtained by raising the price of one product to infinity while decreasing it for the remaining products (detailed discussion in [3]). Obviously, this is not only from a structural, but also from a practical perspective unacceptable. A different source for logical inconsistencies is that, depending on the model configuration, negative values for demand or for prices lead to implausible price sets. In this regard, Zenor [127] describes the tradeoff between logical consistency in the domain of prediction and logical consistency in the domain of optimization for different demand model configurations. As our fifth and last requirement we therefore state that given the demand model, the formulation of our optimization system should account for the above and must not describe an unbounded solution space nor yield negative prices or extreme price sets as optimal solutions.

It is worth to reflect on what is meant by 'extreme price sets': Even though profit maximization is the undisputed long run goal of every retail operation, as a sole objective, it would be misguiding for the short- and mid-term. Retailers are very keen to balance profitability with other objectives such as revenue, unit sales, or market share as well as softer objectives such as price image or assortment attractiveness. This highlights that operational retail pricing is actually a multi-objective problem in which extreme, profit optimal solutions that might involve pricing a product out of the market or losing large amounts of sales or revenue are normally unacceptable. Nonetheless, if the model formulation prevents extreme prices as described above, the price sets produced tend to be useful even under single objective optimization.

For the sake of completeness, we also want to comment on the tractability of the model as optimization literature normally puts emphasis on the topic. Due to the practical approach of this paper, we do not consider this to be a critical characteristic as increased computing power and the existing numerical methods make it nearly irrelevant for practical purposes. This seems counterintuitive at first as we discussed the large scale of the retail problem earlier. Natter et al. [95] even describe in a Marketing Science practice prize report how their industry application effectually works with full enumeration.

4 Demand models for retail pricing

4.1 Overview

In the following we want to review the causal models at the center of the price optimization system which are usually discussed in literature as demand or sales response models. Since we are concerned with the static pricing case, the optimal price set is virtually independent of the demand level. Unless it is treated jointly with questions from inventory management or allocation, non-causal models find little attention in the relevant literature. We therefore

exclude time-series models as well as stochastic models from our review; while these are essential for the dynamic pricing problem they are of negligible importance for our purposes. Even though some of the models used for purely descriptive purposes are of little use for the normative case, others can be very beneficial and are therefore included. Further, most of the models mentioned in the promotional literature are very close to the models needed for our purposes and will therefore be considered.

Various useful taxonomies can be found in literature that organise along dimensions such as dynamic effects, uncertainty handling, individual versus aggregated models, and the model level [74], or intended use of the model (descriptive, predictive, normative), behavioural detail, and consideration of competition [71]. For our purposes, we adapt a frequently used taxonomy and organise the models in two categories: absolute and relative demand models. An absolute demand model, also known as a product-level model, is any model that specifies the price-demand relationship focusing on the individual unit, such as stock keeping unit (SKU) or product or an aggregation thereof such as brand or category, as the level of analysis without (or only secondary) regard to the market environment. The dependent variable is usually an absolute performance figure such as revenue or unit sales. A relative demand model is any model where the share of sales of a unit such as SKU or brand is modeled in relation to a group or aggregate such as category. The dependent variable for our purposes is typically market share.

4.2 Absolute demand models

4.2.1 Model configurations

In their simplest configuration, demand models are strictly linear in the form of the standard regression model as described in (1), where Q is the dependent, absolute performance variable, $X_1, X_2, \ldots, X_n$ are covariates, $a_0, a_1, \ldots, a_n$ are linear parameters, and ε_i is a normally distributed error term. Additional flexibility can be achieved with a nonlinear configuration, yet it usually comes at the cost of increased complexity. A whole range of models offers this flexibility but can be transformed to a linear formulation: The multiplicative model (also known in Economics as Cobb-Douglas Response Function) (2) or its single variable equivalent known as the power function (3). Another important linearizable configuration is the exponential model (4). All of the above can be linearized by logarithmic transformation. The result can be described as a more general form of the strictly linear model discussed above that can be stated as described in (5) where g represents a transforming function of the variables. This kind of configuration is commonly known as the double-log (also linear-in-logs or log-linear, or, if not all covariates are in logarithmic form, as mixed-log) model (6). The semi-logarithmic model is the equivalent with the response variable in its non-logarithmic form. For the sake of completeness, we want to mention models that are non-linear and can not as easily be linearized and are hence intrinsically non-linear (examples in [74, p.76]). While most of the relative sales models presented in Section 4.3 use a nonlinear form, the use of inherently nonlinear absolute models for the purpose of price optimization purposes is rare.

$$\textit{Strictly linear}: \; Q = a_0 + a_1 X_1 + ... + a_n X_n + \varepsilon_i \tag{1}$$

$$\textit{Multiplicative}: \; Q = a_0 X_1^{a_1} X_2^{a_2} ... X_n^{a_n} \varepsilon_i \tag{2}$$

$$\textit{Power}: \; Q = a_0 X_1^{a_1} \varepsilon_i \tag{3}$$

$$\textit{Exponential}: \; Q = a_0 e^{a_1 X_1 + a_2 X_2 + ... + a_n X_n + \varepsilon_i} \tag{4}$$

$$\textit{General}: \; g_0(Q) = a_0 + a_1 g_1(X_1) + a_2 g_2(X_2) + ... + a_n g_n(X_n) + \varepsilon_i \tag{5}$$

$$\textit{Double-log}: \; ln(Q) = a_0 + a_1 ln(X_1) + a_2 ln(X_2) + ... + a_n ln(X_n) + \varepsilon_i \tag{6}$$

4.2.2 Properties of model configurations

A key advantage of linear models is their ease of use and interpretation. Even though linear models can be criticized for being overly simplistic, they tend to provide good local approximations of more complicated formulations. For price optimization purposes, where an optimal solution that is acceptable from a practical point of view must usually be in close proximity of the current price, profit, and revenue combination, this is not a critical limitation. Further, a linear or linearized model can draw from a broad and well developed range of estimation instruments known from regression analysis. This tool set offers well explored solutions for many data properties or challenges seen in retail data and usually allows estimation based on store-level scanner data.

Returns to scale of the variables as determined by the model configuration and the corresponding implications for the (price) elasticities are of great importance for the optimization system. While the strictly linear model is characterized by constant returns to scale, and the exponential model by increasing returns to scale, the multiplicative model offers additional flexibility and can accommodate increasing as well as decreasing returns to scale. The latter appears particularly useful when sales are expected to go towards 0 for large values of price and price decreases are assumed to feature increasing returns to scale. An advantage of the semi-logarithmic configuration is the intuitive interpretation of the dynamics of the resulting model: constant percentage changes in one of the independent variables lead to constant absolute changes of the dependent variable, which in terms of sales resonates with the idea of a saturation limit. In terms of elasticities, this translates to $\epsilon_{X\,i} = \frac{a_i X_i}{a_0 + a_1 X_1 + ... + a_n X_n}$ for the strictly linear formulation while in the double-log form, elasticities are constant and hence the parameter can be interpreted directly as the elasticity so that $\epsilon_{X\,i} = a_i$. For the semi-log model, the elasticities $\epsilon_{X\,i} = \frac{a_i}{a_0 + a_1 ln(X_1) + ... + a_n ln(X_n)}$ mirror the diminishing returns to scale imposed by the log formulation of the parameters that translate into an absolute change of the response variable.

The capability of the variables to assume negative values is a property that is problematic due to the logical inconsistency it implies. This is possible in a strictly linear configuration and will lead to implausible price sets or predictions. Non-linearity in the relevant parameters such as it is the case in a log-log model, and (with minor restrictions) the semi-log model will prevent this.

Interaction between variables can be included in the model formulation and the various model forms offer different degrees of flexibility to do so. This significantly complicates the model even with a small number of variables. As mentioned previously, one of the most critical aspects in a retail pricing context is the consideration of cross price effects. While this is easily feasible for all model forms discussed, the inclusion of additional parameters for competing products will quickly lead to a degrees of freedom problem as we will illustrate in Section 5.

4.2.3 Instances in literature

In terms of covariates, the most simplistic models found in literature take one of the formulations named above and only include own price and the prices of competing products (e. g. [12], [121], [7], [100]).

While promotional effects tend to be a critical influence on retail sales, their consideration in a demand model is not as straightforward as price since promotional efforts can take various formats. Many models consider dummy variables indicating activities normally referred to as display, deal, advertising or feature (e. g. [58], [85], [35], [123]) or several of these (e. g. [129], [85]). In rare cases, interactions between two ([31], [69], [90], [103], [114], [124]) or even three ([96], [68]) of these promotional effects are considered. The inclusion of cross promotion effects between products is quiet rare and only sometimes considered for instances with a small number of products (e. g. [90]). In specialised models, the inclusion of promotions is more elaborate and includes influences such as a maximum deal discount for competing brands [14] or a price-cut ratio [103]. In this context, an area of special interest that has been extensively researched is sales effects of promotions before or after the promotional activity. Blattberg et al. define a deal decay variable that indicates the current week of an n-week multiweek deal [16]. More commonly, pre- or post-promotions lag effects are considered in various ways (e. g. [14], [77] ,[90], [114], [116]).

One of the most proliferated, specialized models for promotions is the SCAN*PRO model [124]. It uses a multiplicative formulation to model weekly store-level brand sales including variables for relative price, dummy variables for feature, display, or both, and indicators for week and store. It has been extensively applied in practice and research alike and has been adapted and expanded in numerous ways (e. g. [31], [35], [43], [67], [114], [115]).

Beyond the effects of promotional activities, there are temporal factors which can have a large impact on retail sales, foremost seasonality. The most proliferated approach is the inclusion of dummy variables that capture seasonality at a suitable level such as one variable indicating season or off-season [16], quarterly variables for each season (e. g. [58], [123]), or more granular solutions such as month and day (e. g. [7], [46]). In more sophisticated formulations, trigonometric terms are included [95]. Depending on the nature of the product, temperature can serve as a proxy for season or weather in general can be included in the formulation (e. g. [129]).

4.3 Relative demand models

4.3.1 Theoretical background

Relative sales models are often discussed as brand share, or market share models, even though the models are more versatile than this. While it is feasible to use any of the absolute model formulations discussed above to model market share instead of an absolute performance measure, the main structural shortcoming to this approach is a lack of logical consistency: It is desirable that market share is bound between 0 and 1 and that the sum of all shares will add up to 1 (e. g. [53, p.121] or [71, p.171]). In the following, we will discuss alternative approaches that naturally assure this integrity. We want to briefly highlight two essential concepts that motivate the better part of the models presented and that will serve as a basis for our discussion.

Traditionally, market share is considered to be an aggregated quantity of a product's or brand's individual sales in relation to the overall category sales. Bell et al. [10] stated in their Market Share Theorem that the market share S of a product i equals its attraction A

relative to the sum of the attractions of all products such that $S_i = \frac{A_i}{\sum_{j=1}^{i} A_j}$. The important difference here is that the market share of an individual brand or product does not only depend on its own variables (e. g. price or marketing effort) but is also directly influenced by the share of the competing brands or products and hence their attraction. While this idea can be motivated in different ways, it has become a key differentiator to conventional modeling concepts of brand share.

Another critical concept that motivates an entire class of models is to consider individual buying decisions of the consumers and derive the product's market share from its probability of being purchased. Models that build on this concept are generally referred to as (Discrete) Choice Models or Random Utility Maximization models. Here, customer's purchasing decisions among different alternatives are modeled rather than the sales of an individual product. The model is a straight forward extension of the logistic regression. At its centre is the decomposition of the utility of an option (i.e. a product) as the sum of a deterministic component u_i and a random component ε_i, so that the total utility can be expressed as $U_i = u_i + \varepsilon_i$. While the deterministic component expresses a utility that is perceived identically by all buyers, the random component represents customer heterogeneity. Therefore, realized utility between two buyers may be different even though the expected utility is the same. This can be interpreted as the heterogeneity of preferences across customers or as the unobservable factors affecting the utility of the product for the individual.

4.3.2 Model configurations

The most relevant models to operationalise market share draw from the ideas above. In terms of functional form, there are many different ways to formulate attraction A. A prominent example here is the Multiplicative Competitive Interaction (MCI) model, popularized by Cooper et al. [33], which formulates attraction as described in (7). Popular extension of this model explicitly include cross effects and are known as the Differential-Effects (or Extended) MCI and the fully extended MCI model [25].

A more contemporary approach draws from choice theory. The most widely used choice model is also the most proliferated relative sales model: The Mutinomial Logit (MNL) model, as formulated in (8). The random component follows a Gumbel distribution, so that its cumulative distribution is $F(\varepsilon_i) = e^{-e^{-\varepsilon_i}}$, and assumes independence of the errors. This error formulation is the key differentiator to the less frequently used Probit model that relies on a normal distribution of the error term and allows covariance between error terms to be non zero. A popular variation of the above is the Nested Logit Model which allows choices to be partitioned into subsets. Further, the popular Mixed Logit model relaxes some crucial restrictions by combining aspects of Probit and MNL. Various extensions of these ideas have been proposed.

$$MCI: \ A_i = exp(a_0) X_1^{a_1} ... X_n^{a_n} \varepsilon_i \tag{7}$$

$$MNL: \ A_i = exp(a_0 + a_1 X_1 + ... + a_n X_n + \varepsilon_i) \tag{8}$$

4.3.3 Properties of model configurations

In terms of returns to scale, the formulations are similar to their absolute counterparts: Since choice models have an exponential form, they do not allow for decreasing returns to scale. The multiplicative form of the MCI offers more flexibility in this regard.

A general disadvantage of relative sales models is that parameter estimates are generally harder to interpret. Also elasticities can not be used as commonly defined since the dependent

variable is a share value between 0 and 1. A popular alternative are Quasi-Elasticities which relate a change in market share to a change in price $e_i = X\frac{\partial s}{\partial X}$, so that for a simple effect configuration we can formulate the elasticities as $e_i = \beta_l(1-m)$ for the multiplicative case and $e_i = \beta_l(1-m)x_{lj}$ for the MNL. It is clear that the elasticity in both formulations is dependent on the product's market share. A desirable property in this context is that the market share elasticity approaches 0 as share goes to 1 since any additional price decrease will only yield a small percentage increase in market share. This property holds for all model formulations presented above. Further, the formulation visualises that, as a fundamental difference between the two formulations, the elasticity in the MNL depends on the current price level.

While the vast majority of absolute models discussed were linear or linearizable, all relative models named here are inherently non-linear and generally not linearizable with standard transformation strategies. Approaches for linearization exist but are not without drawbacks (e. g. [92], [71, p.176], [33, p.144]). Hence Maximum Likelihood based estimation methods are normally needed. The complexity that comes with these methods is the reason why these models were available but were not as popular as linear or linearizable models. Only the advent of simulation-based estimation methods and the increase in computing power have facilitated their application and proliferation.

In the context of estimation, it is crucial to discuss the data used to calibrate these models. Due to the origin of choice models, the data traditionally and most commonly used are panel data on the individual or household level. Guadagni et al. [48] pioneered the use of scanner panel data for the estimation of an MNL and paved the way for this convenient and reliable data collection mechanism that many studies have followed. As we pointed out in Section 3, we deem it essential for practical price optimization purposes that the demand model can be calibrated on store level data. While this is the standard for the absolute models presented above, the models reviewed here can often not be operationalised in this way. The implications for the price optimization problem will be discussed in Section 5.

4.3.4 Instances in literature

The covariates normally included in the formulation of these relative models are similar to those considered in the absolute models introduced previously. In the same manner, the most simple models often only include price (e. g. [78], [62], [46]), or additionally a promotion component (e. g. [29]). The attraction formulation of this model class ensure that interaction effects between products are implicitly considered. However, the level of heterogeneity of cross price effects various considerably. Bultez et al. established the now popular distinction into simple effects model which only allows one parameter a_n per variable n which is equal across products, differential effects model, which accommodates individual parameters per product a_{ni}, and cross-effects model which allows for all possible cross effects between products a_{nij} [25]. Further, since most of the models found in literature are laid out for household level data, household specific variables such as brand loyalty (e. g. [48]) or also size loyalty (e. g. [52], [24]) are often included. Many studies also use choice models to model 'category incidence', meaning whether or not a product in the category was purchased. While this of course deviates from our definition of relative sales models, it is directly intertwined with the discussion at hand. The models used in this sense often feature covariates such as household inventory (e. g. [119], [24]), or rate of category consumption (e. g. [24]).

When it comes to the configuration of the models, the MCI model has been extensively used in earlier work (e. g. [25], [26], [113]) but has lost its popularity to choice models which, apart from being more versatile and in general feature advantageous properties, can rely

on a sound theoretical motivation. Within the class of choice models, the footprint of the Probit model for the problem at hand is comparably small due its complexity when it comes to analytical expressions for its choice probabilities, as well as for the general challenges for its estimation and evaluation. However, we want to highlight the special importance of nested formulations for our problem: One of the most discussed properties of choice models is Independence of Irrelevant Alternatives (IIA), which in our case suggests that if an option is split in equal alternatives that are equivalent for the customer such as the same product with a new UPC or a different color, the combined market share of the split products will increase while the share of the competing products will decrease. Obviously, this conflicts with empirical evidence and managerial wisdom. Many approaches have been proposed to relax this property and most rely on organising the choice alternatives in hierarchies or trees. The Nested Logit model is one way to relax the IIA property and has been used successfully in many situations (e. g. [23], [111]). Moreover, the hierarchical structure imposed here can also be considered consistent with behavioural aspects of the choice process as it is believed that the purchasing decision is organised in hierarchical stages [42].

4.4 Other model forms and dimensions

Beyond the models that are classified and discussed above, there are many other model forms, dimensions, and modeling approaches that can be interesting in the pricing context. In the following, we want to provide some exemplary references for a few that we consider especially relevant.

Reference price: Behavioural aspects can greatly affect sales response and especially the reception of promotional stimuli. It is worth to consider aspects of prospect theory, mental accounting, and, for our purposes most relevant, reference price (review in [82]) in RM systems [118]. Kalyanaram et al. [60] identify empirical generalizations from reference price research while Briesch et al. [20] provide a comparative study of several models. Natter et al. [95] include a reference price component in the demand model of their price optimization system.

Game theory: Competitive effects are intensely studied in the retail context. To analyse the influence of competitive dynamics using the tools of game theory has proved to be very rewarding. In the retail scenario, relevant influences are for one customers that act strategically to promotions and markdowns which can have an impact on their sales response to undiscounted items. For another, manufacturers can react to prices set by the retailer by adjusting purchasing conditions and trade marketing payments for that retailer which can directly affect cost and profit. Further, competing retailers can strategically respond to price changes by adjusting their prices which again can have an impact on sales response. Moorthy [88], [89] provides a general overview of models. Even though game theoretical models are expected to have an increasing importance in the price optimization context in the future (e. g. [118]), an example in literature of their explicit use for normative static retail pricing has not come to our attention.

Hedonic pricing: A very traditional approach to pricing is the idea of hedonic pricing where the price of an item is determined as a combination of its attributes. The approach has a long history in economics and has been enjoying popularity for pricing real estate, while generating only little interest in retail (e. g. [79]). However, an interesting approach that is useful for our purposes draws from the same idea: As the level of analysis typically is

either SKU, brand, or category, it can be advantageous to build models that use product characteristics as the basis of analysis. A small number of studies have made use of this concept in the context of retail demand modeling (e. g. [40], [42], [54], [117]).

Models from Economics: Demand modeling is a traditional field of interest of Economics. We want to highlight some models that are rooted in economic theory, and that focus on economic influences such as disposable income, or a firm's advertising expenditure (e. g. [104], [70]). Some of these models are well known and have been used and re-interpreted for decades such as the Translog Model [32], the Generalized Leontief model [37], and most prominently the Almost Ideal Demand System (AIDS) [36], and the Rotterdam model ([9], [112]). These models have been extended, extensively compared (e. g. [6], [8]), and have inspired and influenced other models (e. g. [14]). There are also contributions that adapt and re-interpret them for our purposes (e. g. [34], [3]).

Semi- and non-parametric models: We have focused on parametric models, but semi- and non-parametric approaches exist that allow more flexibility in response modeling. Models based on non-parametric or semi-parametric regression techniques (e. g. [59], [107]), Neural Networks (e. g. [56]), or Support Vector Machines (e. g. [80], [81]) have been used to model promotion and sales response in retail and can also serve as a basis for price optimization.

4.5 Discussion

As we have seen, there is a vast number of models as well as a plethora of modeling options available and naturally a single superior model for our purposes does not exist. Many of the studies using the models proposed have a primarily descriptive scope and do not share our normative focus. Further, the form of the model ultimately used is highly dependent on the aim of the study and the data available. Nonetheless, researchers obviously try to construct the model that yields the best performance, but even the criteria to assess this performance are disputable and can rely on various quantification concepts of model fit and forecasting accuracy. This is further complicated by the researcher's discretion in modeling and data processing.

We want to discuss functional form first. The question what is popular in literature is easier to assess than what is appropriate for our purposes. Due to its limited flexibility, there is a general criticism in literature for the strictly linear model (e. g. [75], [58]). Depending on era and scope, all other linearizable absolute configurations have been popular choices for modeling retail demand. Wildt [123] states that the most often found additive models are log-log and semi-log. Bitran et al. [13] describe the exponential configuration as commonly used to model demand in retail. Steiner et al. [107] state that the multiplicative, semi- and double-log models are the most popular choices to include nonlinearity in price response models. In a 1988 meta study by Tellis [110] of 424 models from 42 studies between 1960 and 1985, the two most common functional forms were additive and multiplicative models (including exponential and semi-log models). While attraction style formulation certainly has been used for decades, the rise in popularity of choice modeling facilitated by better estimation techniques has fairly recently shifted the field, so that choice models and foremost the MNL have become ubiquitous. Gupta [51] notes that regression models are often used to analyze store level data while the MNL is the preferred model for brand choice.

Many studies address which functional forms deliver the best results in their particular scenario (e. g. [119], [58]), or even focus on the comparison of model formulations under different objectives and applying different criteria (e. g. [44], [21], [57]). Most interesting for

our problem are contributions that evaluate functional forms of models according to their capabilities of determining price elasticities: Bolton [17] studies differences between a linear, multiplicative and exponential formulation in own and cross price elasticity estimates. She concludes that results are very similar, yet overstatement is lowest for the multiplicative form while there can be significant differences across stores. Tellis [110] can not find statistical significant influence of the functional form on price elasticities. A direct comparison with relative models is difficult due to the definition of price elasticity outlined earlier. Hanssens et al. [53, p. 242] provide an overview of comparative studies that analyse functional forms of market share models.

We can summarize that the two groups presented above do not only show some fundamental differences in the formulation of their analytical objective but also differ in properties that are likely to affect the optimization system. While in the more traditional approach of modeling sales in absolute terms we can rely on easy accessible instruments for estimation, the modern, and theoretically more appealing concept of modeling relative market share comes with increased complexity in model formulation and estimation. We further see that the choice models introduced and foremost the MNL are ubiquitous due to a solid theoretical base. The MCI model as an alternative relative formulation has virtually vanished from literature and can be disregarded going forward. We now want to discuss the modeling concepts reviewed in light of the five criteria for applicability defined earlier in Section 3.

5 Demand models in the Revenue Management context

In Section 3, our first criterion from a data perspective was the **reliance on operational (store-level) data only**. While this is common for absolute models, we see that the majority of choice models in literature rely on panel data. Apart from the obvious challenge that these models often include covariates that can not be operationalised with aggregate data, other challenges are more severe. To discuss this further, we want to briefly revisit the original theoretical motivation of choice modeling: With its focus on the decision of an individual as probability of choice given the attributes of an individual, the application to model market share given the attributes of a product seems to depart considerably from the original intention. Therefore two major re-interpretations are necessary. First, we need to rely on the characteristics of the choice alternatives rather than the attributes of the individuals. The theoretical groundwork for this is provided with the conditional logit model. Second, we need to re-interpret choice probabilities as market shares which from a theoretical point of view is the bigger problem. Studies that rely on household level data do not need to make this assumption as choice probability of the household within the category can be modeled, which is not possible when using store level data. Choice probabilities are then often simply re-interpreted as market shares: e. g. Guadagni [48] models choice behaviour for a coffee category based on panel scanner data and derives market share by assuming that "for a given population the average probability of choosing an alternative is the expected share of choices for that alternative".

However, to make such an assumption based on store level data is not without problems: Hardie et al. [54] name as main concerns the extreme assumptions about buying behaviour, the unrealistic assumption of an underlying multinomial process, and the problems with the random error component when using maximum likelihood estimation. A considerable body of literature exists though that contrasts and attempts to consolidate store-level and household-level estimation (e. g. [5], [51], [62], [128]). From a practical point of view, often little or no consequences are found when moving from household level to aggregated data

(e. g. [5], [2]). The discussion is directly intertwined with the question whether the inclusion of customer heterogeneity is beneficial and many studies present approaches to recover said heterogeneity from aggregated data (e. g. [28], [11], [62]).

In this context, a different, for its theoretical groundwork highly interesting research stream is a fairly recent development in Econometrics. To model a ratio, Papke et al. [97] introduced a generalized linear model named the fractional logit model that relies on a binomial distribution and a logit link function. This idea has been extended to the multinomial case (e. g. [64], [91], [106], [125]) and provides a different approach to the problem at hand. An (explicit) application to the retail price optimization problem has not come to our attention yet.

In the following we would like to discuss two criteria in conjunction. For the demand model we argued that the **inclusion of cross price effects** is imperative as intra assortment product dependency is one of the defining properties of the retail problem. From a data perspective, we also required the **suitability for industry size problems**. While even the estimation of a simplistic model can already be considered a challenge if done on a realistic scale, the inclusion of cross price effects considerably complicates this problem.

We want to briefly illustrate the above. If we were to ignore any cross price effects in a demand model formulation, we could either include a single, unified price parameter or an individual price parameter for each of N products. The obvious advantage of an attraction based over an absolute formulation is that even with such a simplistic model, interdependencies between products are implicitly captured. In this context, Mantrala et al. [78] point out the inferiority of an absolute log-log model for their three product scenario, because it would involve nine price parameters in comparison to one for the MNL. However, also in choice models, it can be advantageous to explicitly include cross effects in the model. Carpenter at al. [26] argue that an explicit inclusion can help dealing with the IIA property. If cross price effects are explicitly included, the number of parameters increases rapidly in any model configuration: In the context of Cooper's MCI mode, we already briefly discussed the different levels of cross price effects potentially considered. The total number of parameters needed depends on the exact model formulation but if we only consider a product specific, (intersect or attraction like) parameter and the price parameters, there would be $N + 1$ parameters to estimate for the simple effects model, and $N + N$ parameters if we include product-specific price parameters like it is the case in a differential effects model. Accounting for cross price effects in the spirit of a fully extended model, we already need to determine $N + N + 0.5\,N\,(N - 1)$ parameters if we consider symmetric cross price elasticities, or even $N + N + N\,(N - 1) = N + N^2$ parameters if we include asymmetric cross price effects.

Since even a smaller category has 50 active SKUs at any given time, this would already sum up to 2550 parameters to be estimated. With the number of parameters increasing quadratically with N, and considering the large product counts of a realistic retail environment, a degrees of freedom problem is easily visible. Several studies consider different levels of cross effects in their model construction (e. g. [46], [100]). The estimatability when done with the standard, well known methods, depends on the level of parameters included in the model formulation and the size of the problem as determined by the data available. It is obvious that in most applied situations, normative models can not 'afford' to include effects at this level of granularity without retreating to more advanced estimation techniques. As a result, many academic papers rest on assumptions that can not be upheld when considering the reality of the retail environment. It is common to reduce complexity by only considering 5 or 10, rather than 50 or 100 items per category. Even when working with real data, researchers normally prune or aggregate data and retreat to pricing on brand rather than product level which certainly is not possible for any practical application. The popular

practice of focussing on the strongest products in the assortment is especially unfortunate from a practical perspective: Natter et al. [95] describe how their decision support system is especially helpful for low selling products as the retailer uses discretional pricing for the most popular items.

Fortunately, next to the high amount of products, the retail environment typically also features a high volume of transactions, an assortment that spans multiple categories, and often a large amount of stores. This information can be used on different levels to facilitate the estimation process and add stability to parameter estimates. Many contributions make use of this by using pooling (e. g. [31]) or Bayesian methods (e. g. [14], [85]) within the assortment context.

The aspect of multiple stores warrants special attention. On the one hand these data can be used as additional information to strengthen estimates (e. g. [78], [105]). On the other hand, store heterogeneity can be responsible for variance and many studies account for it in their model (e. g. [87], [7]). In regards to pricing, this also implies that store specific pricing can be beneficial (e. g. [78], [87]). A popular approach in theory and practice alike is the creation of clusters for pricing or estimation purposes usually discussed as zone pricing. The determination and utilization of such clusters has attracted much research attention (e. g. [30], [78]).

Within the context of the data model, we also formulated the requirement that the model should have the **potential to accommodate typical retail data conditions**. While this is a field that is intensely cared for in the forecasting literature, it is neglected in a retail pricing context. We will only discuss one example here: Dealing with censored demand is a prominent topic in forecasting literature and has also been popular in RM [50]. Vulcano et al. [120] propose a demand untruncation method based on store level retail data while Kok et al. [65] take a similar approach in an retail assortment planning context. The topic has not been discussed in a normative retail pricing scenario yet.

In terms of our requirements for the optimization model, it is most important that our model produces **sensible solutions when used for optimization**. Unfortunately, all demand models discussed above encounter problems when used in their simple form for optimization purposes as they yield infinite prices [3]. In literature, three different approaches can be found to avoid this issue:

1. A common solution is to impose restrictions on the size of the price changes (e. g. [100]). While retailers usually work with pre-defined maximum price changes, the outcome of the optimization would entirely rely on this arbitrarily defined restriction which is unsatisfying and unacceptable from a practical as well as from a theoretical perspective.
2. A more appealing alternative for attraction-style models is to define an 'external good', which represents the customer's no-buy option or the decision to buy outside of the category (e. g. [78]). While this is a seamless continuation of the theoretical reasoning behind the choice models introduced, the estimation in most empirical situations becomes difficult given our data requirements. Customers shopping but not buying in a category are not captured in scanner data. Moreover, the determination of quantity and marketing-mix of the external good is not obvious.
3. A better approach to the problem is to explicitly include the category purchase incidence into the model. Often, a second choice model is included to determine purchase probability of the category which will nest the brand choice model (e. g. [24], [119]). Again, operationalisation of the model will be difficult given our data requirements as some sort of absolute reference of category size will have to be determined that will not be available in store level data: e.g Mantrala et al. [78] treat the maximum weekly sales of the product

observed in a store as the weekly market potential, while Vilcassim et al. [119] utilise a household's number of visits to the store. A solution that can be estimated without any compromises is to chose an absolute model for category incidence so that category sales is modeled in dependence of the price level of the category (e. g. [3], [46], [108]).

Apart from solving the problem described above, dividing the demand model in this way has some additional advantages. It corresponds well with our behavioural understanding of the demand process as it is assumed that the decision maker first chooses whether or not to shop in a store or a category and then subsequently makes a choice between the available options. Further, Little et al. [76] formulate the idea of a two stage theory of price setting: once the customer is in the store, he will maximize his utility (short-run price response). However, utility level becomes a policy parameter that determines the long run attractiveness of the store. Further, a combination of models allows the modeler to combine advantages of relative models with those of absolute models such as accounting for seasonal or cyclical components in the absolute model without having to adapt the brand choice model. Such a configuration can also help to reduce multicollinearity.

We also mentioned the chance of extreme or negative prices in our requirements. The problem of optimal prices possibly being negative only arises if a linear formulation is used and can be avoided with practically any other formulation allowing non-linear returns of scale. The occurrence of other extreme prices can usually be attributed to unstable or false parameter estimates, usually with incorrect sign and of unreasonable size, due to data issues. Given a meaningful formulation and acceptable fit of the model, the optimal price set usually does not accommodate extreme prices.

6 Conclusion

We have shown that the static retail price optimization problem and its dynamic counterpart are discussed very differently: While for the latter, an interdisciplinary and application oriented discussion is entertained within the scope of RM, a similar discourse does currently not exist for the former. Instead, the primary discussion of the problem is taking place in the Marketing and Retailing literature, where a long history of very different, and mostly very well explored, demand modeling approaches are available to form the theoretical centre of the static retail price optimization problem. Naturally, said literature is primarily concerned with descriptive studies foremost focused on price elasticities, not with the normative questions and the implications of the systemic context derived from data, optimization and implementation. The body of literature addressing these questions is currently very small. We have seen that when these models are analysed with an applied, and integrative view, they collide with the practical requirements imposed by the environment of the price optimization system. Accordingly, there is a wide variety of areas that show great potential for future research. We believe that any sensible contribution on the topic that subscribes to such an integrative view in the spirit of RM is worthwhile. However, we want to highlight a few points of interest for which we see the most pressing need for research:

(Realistic) data conditions: We saw that every relevant paper engaging in empirical evaluation does so with an unrealistically small number of brands or products while using models or estimation techniques that are not fit for larger instances. Further, studies explicitly dedicated to analyzing the effects of the typical data conditions regularly studied in time-series analysis do not exist. We can encourage any effort that addresses the above

by advancing estimation methods and studying the impact of these conditions on the price optimization system, its results and its effectiveness.

Data type, processing and organisation: The majority of studies rely on panel data even though store level data is much more relevant for retailers. Further, there is large discretion in the data treatment process that precedes any empirical evaluation. No unified data processing standard has been suggested yet. In their survey, Levy et al. [72] point to the connected question: "How do retailers group items into categories? What is the best way to categorize?". This is a promising area for future research, which undoubtedly is also very relevant from an RM perspective if linked back to the implications on the system. We can encourage any effort that helps making models accessible for use with store level data. Further, we consider the questions how data preprocessing methods influence the effectiveness and optimality of the pricing system especially pressing. Research in this area should pave the way towards a unified data processing standard for retail price optimization systems.

Operationalisation of choice models: According to van Ryzin [118], a shift from product centred models to choice models is needed. Choice modeling has already come a long way in the last decades. However, a lot of questions concerning the operationalisation of choice models in the retail environment remain unanswered such as the implications of the theoretical compromises made when estimating given store level data, including the assumption of discrete choice, and the handling of particularly large choice sets. Further, pretty much all efforts concerning this rely on the choice set as the critical base for modeling. Linking the usually data driven choice set determination to behavioural theory remains unexplored in the pricing context. We therefore believe that contributions fostering the theoretical base of choice modeling in a price optimization context will have great impact.

Multi-objectivity of the optimization problem: We want to reiterate that this is an area where literature is especially scarce. As the optimization problem as such is comparably simple and not as versatile as its dynamic equivalent, it might appear not as attractive from a purely academic perspective. However, a very intriguing aspect is the multi-facetted and conflicting objectives of retail pricing which goes beyond the scope of the original RM problem. Next to revenue and profit, retailers are often interested in preserving price image, market share or in keeping a competitive price level. While Levy et al. [72] suggest in their review to consider the question "How do these conflicting goals affect their customers and their profits?" we would also like to point to the potential to answer the questions usually asked in an OR context, including problem reformulation, efficient solving algorithms, or even in connection with robust optimization and risk management.

The innovations of the past two decades have paved the way for the practical implementation of retail price optimization systems. While the traditional, dynamic RM systems have matured, the static problem is trailing behind in theory and practice alike. We believe that a discussion of the problem in RM literature would encourage cross-disciplinary and practical oriented research that would go a long way in improving this situation.

References

1 Makoto Abe and Kirthi Kalyanam. Store Sales and Panel Purchase Data: Are They Compatible? *Mimeo*, 1995.

2 Greg M. Allenby and Peter E. Rossi. There is no aggregation bias: Why macro logit models work. *Journal of Business & Economic Statistics*, 9(1):1–14, 1991.

3 Eric Anderson and Naufel J. Vilcassim. Structural demand models for retailer category pricing. *London Business School Mimeo*, 2001.

4 Rick L. Andrews and Imran S. Currim. An experimental investigation of scanner data preparation strategies for consumer choice models. *International Journal of Research in Marketing*, 22(3):319–331, September 2005.

5 Rick L. Andrews, Imran S. Currim, and Peter S.H. Leeflang. A Comparison of Sales Response Predictions From Demand Models Applied to Store-Level versus Panel Data. *Journal of Business and Economic Statistics*, 29(2):319–326, April 2011.

6 Clifford L.F. Attfield. A Comparison of the Translog and Almost Ideal Demand Models. *Mimeo University of Bristol*, 564(4), 2004.

7 George Baltas. Modelling category demand in retail chains. *Journal of the Operational Research Society*, 56(11):1258–1264, March 2005.

8 William A. Barnett and Ousmane Seck. Rotterdam model versus almost ideal demand system: will the best specification please stand up? *Journal of Applied Econometrics*, 23(6):795–824, 2008.

9 AP P Barten. Evidence on the Slutsky conditions for demand equations. *The Review of Economics and Statistics,*, 49(1):77–84, 1967.

10 David E. Bell, Ralph L. Keeney, and John D.C. Little. A market share theorem. *Journal of Marketing Research*, 12(2):136–141, 1975.

11 David Besanko, Sachin Gupta, and Dipak C. Jain. Logit Demand Estimation Under Competitive Pricing Behavior: An Equilibrium Framework. *Management Science*, 44(11):1533–1547, November 1998.

12 Chitrabhanu Bhattacharya and Leonard M. Lodish. An advertising evaluation system for retailers. *Journal of Retailing and Consumer Services*, 1(2):90–100, 1994.

13 Gabriel R. Bitran and Rene Caldentey. An overview of pricing models for revenue management. *Manufacturing & Service Operations Management*, 5(3):203–229, 2003.

14 Robert C. Blattberg and Edward I. George. Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association*, 86(414):304–315, 1991.

15 Robert C. Blattberg and Scott A. Neslin. Sales promotion: The long and the short of it. *Marketing Letters*, 1(1):81–97, 1989.

16 Robert C. Blattberg and Kenneth J. Wisniewski. Price-induced patterns of competition. *Marketing Science*, 8(4):291–309, 1989.

17 Ruth N. Bolton. The Robustness Of Retail-Level Price Elasticity Estimates. *Journal of Retailing*, 65(2):193–219, 1989.

18 Ruth N. Bolton and Venkatesh Shankar. An empirically derived taxonomy of retailer pricing and promotion strategies. *Journal of Retailing*, 79(4):213–224, January 2003.

19 Richard A. Briesch, William R. Dillon, and Robert C. Blattberg. Treating Zero Brand Sales Observations in Choice Model Estimation: Consequences and Potential Remedies. *Journal of Marketing Research*, 45(5):618–632, October 2008.

20 Richard A. Briesch, Lakshman Krishnamurthi, Tridib Mazumdar, and S.P. Raj. A Comparative Analysis of Reference Price Models. *Journal of Consumer Research*, 24(2):202–214, 1997.

21 Roderick J. Brodie and Cornelis Kluyver. Attraction Versus Linear and Multiplicative Market Share Models: An Empirical Evaluation. *Journal of Marketing Research*, 21(May):194–201, 1984.

22 Bart J. Bronnenberg, Michael W. Kruger, and Carl F. Mela. The IRI marketing data set. *Marketing Science*, 27(4):745–748, 2008.

23 Patrick G. Buckley. Nested multinomial logit analysis of scanner data for a hierarchical choice model. *Journal of Business Research*, 17(2):133–154, 1988.

24 Randolph E. Bucklin and James M. Lattin. A model of product category competition among grocery retailers. *Journal of Retailing*, 68(3):271–294, 1992.

25 Alain V. Bultez and Philippe A. Naert. Consistent sum-constrained models. *Journal of the American Statistical Association*, 70(351):529–535, 1975.

26 Gregory S. Carpenter, Lee G. Cooper, Dominique M. Hanssens, and David F. Midgley. Modeling asymmetric competition. *Marketing Science*, 7(4):393–412, 1988.

27 Wen-Chyuan Chiang, Jason C. H. Chen, and Xiaojing Xu. An overview of research on revenue management: current issues and future research. *International Journal of Revenue Management*, 1(1), 2007.

28 Pradeep K. Chintagunta. Endogeneity and Heterogeneity in a Probit Demand Model: Estimation Using Aggregate Data. *Marketing Science*, 20(4):442–456, 2001.

29 Pradeep K. Chintagunta. Investigating Category Pricing Behavior at a Retail Chain. *Journal of Marketing Research*, 39(2):141–154, May 2002.

30 Pradeep K. Chintagunta, Jean-Pierre Dubé, and Vishal Singh. Balancing profitability and customer welfare in a supermarket chain. *Quantitative Marketing and Economics*, 1(1):111–147, 2003.

31 Markus Christen, Sachin Gupta, John C. Porter, Richard Staelin, and Dick R. Wittink. Using Market-Level Data to Understand Promotion Effects in a Nonlinear Model. *Journal of Marketing Research*, 34(3):322, August 1997.

32 Laurits R. Christensen and William Greene. Economies of scale in US electric power generation. *The Journal of Political Economy*, 84(4):655–676, 1976.

33 Lee G. Cooper and Masao Nakanishi. *Market-share analysis: evaluating competitive marketing effectiveness*. Springer, 1988.

34 Ronald W. Cotterill, William P. Putsis Jr., and Ravi Dhar. Assessing the competitive interaction between private labels and national brands. *Journal of Business*, 73(1):109, 2000.

35 Peter J. Danaher, Andre Bonfrer, and Sanjay Dhar. The effect of competitive advertising interference on sales for packaged goods. *Journal of Marketing Research*, XLV(April):211–225, 2008.

36 Angus Deaton and John Muellbauer. An almost ideal demand system. *The American Economic Review*, 70(3):312–326, 1980.

37 Erwin Diewert. An application of the Shephard duality theorem: a generalized Leontief production function. *The Journal of Political Economy*, 79(3):481–507, 1971.

38 Suresh Divakar, Brian T. Ratchford, and Venkatesh Shankar. CHAN4CAST: A Multichannel Multiregion Forecasting Model for Consumer Packaged Goods. *Marketing Science*, 24(3):334–350, July 2005.

39 Wedad Elmaghraby and Pinar Keskinocak. Dynamic pricing in the presence of inventory considerations: research overview, current practices, and future directions. *IEEE Engineering Management Review*, 31(4):47–47, 2003.

40 Peter S. Fader and Bruce G. S. Hardie. Modeling Consumer Choice among SKUs. *Journal of Marketing Research*, 33(4):442, November 1996.

41 Juan-Carlos Ferrer and Diego Fuentes. A system design to bridge the gap between the theory and practice of retail revenue management. *International Journal of Revenue Management*, 5(2):261–275, 2011.

42 Eijte W. Foekens, Peter S.H. Leeflang, and Dick R. Wittink. Hierarchical versus other market share models for markets with many items. *International Journal of Research in Marketing*, 14:359–378, 1997.

43 Eijte W. Foekens, Peter S.H. Leeflang, and Dick R. Wittink. Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometrics*, 89(1-2):249–268, 1998.

44 Avijit Ghosh and Robert Shoemaker. A Comparison of Market Share Models and Estimation Procedures. *Journal of Marketing Research*, XXI(May):202–211, 1984.

45 Jochen Goensch, Robert Klein, and Claudius Steinhardt. Dynamic Pricingâ€“State-of-the-Art. *Zeitschrift für Betriebswirtschaft*, (3):1–40, 2009.

46 Óscar González-Benito, María Pilar Martínez-Ruiz, and Alejandro Mollá-Descals. Retail pricing decisions and product category competitive structure. *Decision Support Systems*, 49(1):110–119, April 2010.

47 Dhruv Grewal, Kusum L. Ailawadi, Dinesh Gauri, Kevin Hall, Praveen Kopalle, and Jane R. Robertson. Innovations in Retail Pricing and Promotions. *Journal of Retailing*, 87(1):43–52, July 2011.

48 Peter M. Guadagni and John D.C. Little. A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2(3):203–238, 1983.

49 Joseph Guiltinan. Progress and Challenges in Product Line Pricing. *Journal of Product Innovation Management*, 28:744–756, April 2011.

50 Peng Guo, Baichun Xiao, and Jun Li. Unconstraining Methods in Revenue Management Systems: Research Overview and Prospects. *Advances in Operations Research*, 2012:1–23, 2012.

51 Sachin Gupta, Pradeep K. Chintagunta, Anil Kaul, and Dick R. Wittink. Do Household Scanner Data Provide Representative Inferences from Brand Choices: A Comparison with Store Data. *Journal of Marketing Research*, 33(4):383, November 1996.

52 Sunil Gupta. Impact of Sales Promotions on When, What, and How Much to Buy. *Journal of Marketing Research (JMR)*, 25(4):342–355, 1988.

53 Dominique M. Hanssens, Leonard J. Parsons, and Randall L. Schultz. *Market Response Models: Econometric and Time Series Analysis*. Springer, 2001.

54 Bruce G. S. Hardie, Leonard M. Lodish, Peter S. Fader, Alistair P. Sutcliffe, and William T. Kirk. Attribute-based Market Share Models: Methodological Development and Managerial Applications. *London Business School Mimeo*, 1998.

55 Stephen J. Hoch, Byung-Do Kim, Alan L. Montgomery, and Peter E. Rossi. Determinants of Store-Level Price Elasticity. *Journal of Marketing Research*, 32(1):17, February 1995.

56 Harald Hruschka. Relevance of functional flexibility for heterogeneous sales response models - A comparison of parametric and semi-nonparametric models. *European Journal of Operational Research*, 174(2):1009–1020, October 2006.

57 Dipak C. Jain and Naufel J. Vilcassim. Testing functional forms of market share models using the Box-Cox transformation and the Lagrange multiplier approach. *International Journal of Research in Marketing*, 6:95–107, 1989.

58 Kirthi Kalyanam. Pricing decisions under demand uncertainty: A Bayesian mixture model approach. *Marketing Science*, 15(3):207–221, 1996.

59 Kirthi Kalyanam and Thomas S. Shively. Estimating Irregular Pricing Effects: A Stochastic Spline Regression Approach. *Journal of Marketing Research*, 35(1):16–29, 1998.

60 Gurumurthy Kalyanaram and Russell S. Winer. Empirical generalizations from reference price research. *Marketing Science*, 14(3):161–169, 1995.

61 University of Chicago Booth School of Business Kilts Center for Marketing. Dominick's Data Manual. 2013.

62 Byung-do Kim, Robert C. Blattberg, and Peter E Rossi. Modeling the Distribution of Price Sensitivity and Implications for Optimal Retail Pricing. *Journal of Business*, 13(3):291–303, 1995.

63 Robert Klein. *Revenue Management*. Springer, Berlin, Heidelberg, 1. edition, 2008.

64 Steven F. Koch. Fractional multinomial response models with an application to expenditure shares. *Mimeo University of Pretoria*, 2010-21(October), 2010.

65 Gürhan A. Kök, Marshall L. Fisher, and Ramnath Vaidyanathan. Assortment Planning: Review of Literature and Industry Practice. *Retail Supply Chain Management*, 122:99–153, 2009.

66 Praveen K. Kopalle. Modeling Retail Phenomena. *Journal of Retailing*, 86(2):117–124, June 2010.

67 Praveen K. Kopalle, Carl F. Mela, and Lawrence Marsh. The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications. *Marketing Science*, 18(3):317–332, 1999.

68 V Kumar and Robert P. Leone. Measuring the effect of retail store promotions on brand and store substitution. *Journal of Marketing Research*, 25(2):178–185, 1988.

69 V Kumar and Arun Pereira. Explaining the Variation in Short-Term Sales Response to Retail Price Promotions. *Journal of the Academy of Marketing Science*, 23(3):155–169, June 1995.

70 Jean-Jacques Lambin. Measuring the profitability of advertising: An empirical study. *Journal of Industrial Economics*, 18(2):86–103, November 1969.

71 Peter S.H. Leeflang, Dick R. Wittink, Michel Wedel, and Philippe A. Naert. *Building models for marketing decisions*. Kluwer Academic Publishers, Boston, 2000.

72 Michael R. Levy, Dhruv Grewal, Praveen Kopalle, and James Hess. Emerging trends in retail pricing practice: implications for research. *Journal of Retailing*, 80(3):xiii–xxi, 2004.

73 Warren H. Lieberman. Revenue management trends and opportunities. *Journal of Revenue and Pricing Management*, 3(1):91–99, 2004.

74 Garry L. Lilien and Philip Kotler. *Marketing Decision Making - A Model-Building Approach*. 1983.

75 John D.C. Little. *Models and managers: The concept of a decision calculus*. PhD thesis, 1970.

76 John D.C. Little and Jeremy F. Shapiro. A theory for pricing nonfeatured products in supermarkets. *Journal of Business*, 53(3):199–209, 1980.

77 Sandrine Macé and Scott A. Neslin. The determinants of pre-and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research*, 41(3):339–350, 2004.

78 Murali K. Mantrala, P.B. Seetharaman, Rajeeve Kaul, Srinath Gopalakrishna, and Antonie Stam. Optimal Pricing Strategies for an Automotive Aftermarket Retailer. *Journal of Marketing Research*, 43(4):588–604, November 2006.

79 Josué Martínez-Garmendia. Application of hedonic price modeling to consumer packaged goods using store scanner data. *Journal of Business Research*, 63(7):690–696, July 2010.

80 María Pilar Martínez-Ruiz. Evaluating temporary retail price discounts using semiparametric regression. *Journal of Product and Brand Management*, 15(1):73–80, 2006.

81 María Pilar Martínez-Ruiz, José Luis Rojo-Álvarez, and Francisco Javier Gimeno-Blanes. Evaluation of Promotional and Cross-Promotional Effects Using Support Vector Machine Semiparametric Regression. *Systems Engineering Procedia*, 1:465–472, January 2011.

82 Tridib Mazumdar, S.P. Raj, and Indrajit Sinha. Reference Price Research: Review and Propositions. *Journal of Marketing*, 69(4):84–102, October 2005.

83 Jeffrey McGill and Garrett J. van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 32(2):233–256, 1999.

84 Kent B. Monroe and Albert J. Della Bitta. Models for Pricing Decisions. *Journal of Marketing Research*, 15(3):413–428, 1978.

85 Alan L. Montgomery. Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science*, 16(4):315–337, 1997.

86 Alan L. Montgomery. The implementation challenge of pricing decision support systems for retail managers. *Applied Stochastic Models in Business and Industry - Bridging the*

Gap between Academic Research in Marketing and Practitioners' Concerns, 1(4-5):367–378, 2005.

87 Alan L. Montgomery and Peter Rossi. Estimating price elasticities with theory-based priors. *Journal of Marketing Research*, 36(4):413–423, 1999.

88 Sridhar K. Moorthy. Using game theory to model competition. *Journal of Marketing Research*, 22(3):262–282, 1985.

89 Sridhar K. Moorthy. Competitive marketing strategies: Game-theoretic models. In *Handbooks in Operations Research and Management Science*, volume 5, pages 143–190. 1993.

90 Mark M. Moriarty. Retail promotional effects on intra-and interbrand sales performance. *Journal of Retailing*, 61(3):27–47, 1985.

91 John Mullahy. Multivariate fractional regression estimation of econometric share models. *University of Wisconsin-Madison Mimeo*, WP2011/33, 2010.

92 Masao Nakanishi and Lee G. Cooper. Simplified Estimation Procedures for MCI Models. *Marketing Science*, 1(3):314–322, 1982.

93 Martin Natter, Thomas Reutterer, and Andreas Mild. Dynamic Pricing Support Systems for DIY Retailers - A case study from Austria. *Marketing Intelligence Review*, 1(1):17 –23, 2009.

94 Martin Natter, Thomas Reutterer, Andreas Mild, and Alfred Taudes. Ein sortimentsübergreifendes Entscheidungsunterstützungssystem für dynamische Preis- und Werbeplanung im DIY-Handel. *Marketing Zeitschrift für Forschung und Praxis*, 28(4):260–268, 2006.

95 Martin Natter, Thomas Reutterer, Andreas Mild, and Alfred Taudes. Practice Prize Report: An Assortmentwide Decision-Support System for Dynamic Pricing and Promotion Planning in DIY Retailing. *Marketing Science*, 26(4):576–583, July 2007.

96 Purushottam Papatla and Lakshman Krishnamurthi. Measuring the dynamic effects of promotions on brand choice. *Journal of Marketing Research*, 33(1):20–35, 1996.

97 Leslie E. Papke and Jeffrey M. Wooldridge. Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics*, 11(February):619–632, 1996.

98 Robert L. Phillips. *Pricing and Revenue Optimization*. Stanford University Press, Stanford, 1. edition, 2005.

99 Vithala R. Rao. Pricing research in marketing: The state of the art. *Journal of Business*, 57(1), 1984.

100 David J. Reibstein and Hubert Gatignon. Optimal Product Line Pricing: The Influence of Elasticities and Cross-Elasticities. *Journal of Marketing Research*, 21(3):259, August 1984.

101 P.B. Seetharaman, Siddhartha Chib, Andrew Ainslie, Peter Boatwright, Tat Chan, Sachin Gupta, Nitin Mehta, Vithala R. Rao, and Andrei Strijnev. Models of Multi-Category Choice Behavior. *Marketing Letters*, 16(3-4):239–254, December 2005.

102 Venkatesh Shankar and Ruth N. Bolton. An Empirical Analysis of Determinants of Retailer Pricing Strategy. *Marketing Science*, 23(1):28–49, January 2004.

103 Venkatesh Shankar and Lakshman Krishnamurthi. Relating price sensitivity to retailer promotional variables and pricing policy: An empirical analysis. *Journal of Retailing*, 72(3):249–272, 1996.

104 Jorge M. Silva-Risso, Randolph E. Bucklin, and Donald G. Morrison. A decision support system for planning manufacturers' sales promotion calendars. *Marketing Science*, 18(3):274–300, 1999.

105 Jorge M. Silva-Risso and Irina Ionova. Practice Prize Winner-A Nested Logit Model of Product and Transaction-Type Choice for Planning Automakers' Pricing and Promotions. *Marketing Science*, 27(4):545–566, 2008.

106 Aruna Sivakumar and Chandra Bhat. Fractional Split-Distribution Model for Statewide Commodity-Flow Analysis. *Transportation Research Record*, 1790(1):80–88, January 2002.

107 Winfried J. Steiner, Andreas Brezger, and Christiane Belitz. Flexible estimation of price response functions using retail scanner data. *Journal of Retailing and Consumer Services*, 14(6):383–393, November 2007.

108 Shivaram Subramanian and Hanif D. Sherali. A fractional programming approach for retail category price optimization. *Journal of Global Optimization*, 48(2):263–277, November 2009.

109 Kalyan T. Talluri and Garrett J. van Ryzin. *The Theory and Practice of Revenue Management*. Springer US, New York, 1. edition, 2005.

110 Gerard J. Tellis. The price elasticity of selective demand: A meta-analysis of econometric models of sales. *Journal of Marketing Research*, 25(4):331–341, 1988.

111 Gerard J. Tellis. Tackling the retailer decision maze: Which brand to discount when and why? *Marketing Science*, 14(3):271–299, 1995.

112 Henri Theil. The information approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 33(1):67–87, 1965.

113 Glen L. Urban. A mathematical modeling approach to product line decisions. *Journal of Marketing Research*, 6(1):40–47, 1969.

114 Harald J. van Heerde, Peter S.H. Leeflang, and Dick R. Wittink. The Estimation of Pre- and Postpromotion Dips with Store-Level Scanner Data. *Journal of Marketing Research*, 37(3):383–395, August 2000.

115 Harald J. van Heerde, Peter S.H. Leeflang, and Dick R. Wittink. How Promotions Work: SCAN*PRO-Based Evolutionary Model Building. *Schmalenbach Business Review*, 54(July):198–220, 2002.

116 Harald J. van Heerde, Peter S.H. Leeflang, and Dick R. Wittink. Decomposing the Sales Promotion Bump with Store Data. *Marketing Science*, 23(3):317–334, June 2004.

117 Erjen van Nierop, Dennis Fok, and Philip Hans Franses. Sales models for many items using attribute data. *ERIM Report Series Research in Management*, (ERS-2002-65-MKT), 2002.

118 Garrett J. van Ryzin. Models of demand. *Journal of Revenue & Pricing Management*, 4(2):204–210, 2005.

119 Naufel J. Vilcassim and Pradeep K. Chintagunta. Investigating retailer product category pricing from household scanner panel data. *Journal of Retailing*, 71(2):103–128, 1995.

120 Gustavo Vulcano, Garrett J. van Ryzin, and Richard M. Ratliff. Estimating Primary Demand for Substitutable Products from Sales Transaction Data. *Operations Research*, 60(2):313–334, May 2012.

121 Rockney G. Walters. Assessing the Impact of Retail Price Promotions on Product Substitution, Complementary Purchase, and Interstore Sales Displacement. *Journal of Marketing*, 55(2):17, April 1991.

122 Lawrence R. Weatherford and Samuel E. Bodily. A taxonomy and research overview of perishable-asset revenue management: yield management, overbooking, and pricing. *Operations Research*, 40(5):831–844, 1992.

123 Albert R. Wildt. Estimating models of seasonal market response using dummy variables. *Journal of Marketing Research*, 14(1):34–41, 1977.

124 Dick R. Wittink, M.J. Addona, W.J. Hawkes, and J.C. Porter. SCAN* PRO: The estimation, validation and use of promotional effects based on scanner data. *Internal Paper, Cornell University*, 1988.

125 Xin Ye and Ram M. Pendyala. A Model of Daily Time Use Allocation Using Fractional Logit Methodology. In *Transportation and Traffic Theory. Flow, Dynamics and Human Interaction. 16th International Symposium on Transportation and Traffic Theory*, 2005.

126 Elaine L. Zanutto and Eric T. Bradlow. Data pruning in consumer choice models. *Quantitative Marketing and Economics*, 4(3):267–287, August 2006.

127 Michael J. Zenor. The profit benefits of category management. *Journal of Marketing Research*, 31(2):202–213, 1994.

128 Michael J. Zenor and Rajendra K. Srivastava. Inferring Market Structure With Aggregate Data: A Latent Segment Logit Approach. *Journal of Marketing Research*, 30(3):369–379, 1993.

129 Giulio Zotteria, Matteo Kalchschmidtb, and Federico Caniatoc. The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, 93-94:479–491, January 2005.