Operations Research Proce

Bo Hu
Karl Morasch
Stefan Pickl
Markus Siegle *Edi*

Operatio
Research

Operations Research Proceedings

Bo Hu • Karl Morasch • Stefan Pickl
Markus Siegle
Editors

Operations Research Proceedings 2010

Selected Papers of the
Annual International Conference of the
German Operations Research Society (GOR) at
Universität der Bundeswehr München

September 1–3, 2010

Springer

Editors
Bo Hu
Universität der Bundeswehr
München
Fakultät für Betriebswirtschaft
Werner-Heisenberg-Weg 39
85577 Neubiberg
Germany
bo.hu@unibw.de

Karl Morasch
Universität der Bundeswehr
München
Fakultät für Wirtschafts- und
Organisationswissenschaften
Werner-Heisenberg-Weg 39
85577 Neubiberg
Germany
karl.morasch@unibw.de

Stefan Pickl
Universität der Bundeswehr
München
Fakultät für Informatik
Werner-Heisenberg-Weg 39
85577 Neubiberg
Germany
stefan.pickl@unibw.de

Markus Siegle
Universität der Bundeswehr
München
Fakultät für Informatik
Werner-Heisenberg-Weg 39
85577 Neubiberg
Germany
markus.siegle@unibw.de

ISSN 0721-5924
ISBN 978-3-642-20008-3 e-ISBN 978-3-642-20009-0
DOI 10.1007/978-3-642-20009-0
Springer Heidelberg Dordrecht London New York

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains a selection of refereed papers referring to lectures presented at the symposium "International Conference on Operations Research (OR2010) – Mastering Complexity" held at Universität der Bundeswehr München, September 1–3, 2010. This international conference took place under the auspices of the German Operations Research Society (GOR). It was organized for the first time in partnership with the Italian Operational Research Society (AIRO) which was responsible for the organizations of three topics.

The symposium had about 750 participants from more than 48 countries all over the world; more than 30 colleagues came from Italy and 127 researchers registered as PhD-candidate; this category was offered successfully for the first time. The conference attracted academics and practitioners working in various fields of Operations Research and provided them with the most recent advances in Operations Research and related areas to the general topic "Mastering Complexity".

Complexity is a natural component of the globalization process: Financial markets, traffic systems, network topologies, energy resource management and so forth are all characterized by complex behavior and economic interdependencies. Operations Research is one of the key instruments to model, simulate and analyze such systems, and to answer questions concerning not only efficiency but also safety & security. In fact, gaining optimal solutions, suitable heuristics and efficient procedures for these applications are some of the challenges which were discussed at the international OR2010 and are documented in that volume.

At the conference, 439 regular talks, 2 plenary keynotes, 15 invited semi-plenary talks and 7 invited talks by award recipients were given. Out of those contributions accepted for presentation at the conference, 171 extended abstracts were submitted for the post-proceedings, of which 104 were finally accepted –after a thorough peer-reviewing process– for publication in the present volume. Due to a limited number of pages available for the proceedings volume, the length of each article had to be restricted.

We would like to thank the GOR-board for the abundant collaboration, which we always found to be helpful and fruitful. We are grateful to all the anonymous reviewers, who were asked to review one or more submitted papers. Many thanks to our Italian colleagues Grazia Speranza, Paola Zuddas and Renato de Leone for their cooperation and enormous efforts. We have to say special thanks to Zafer-Korcan Görgülü who invested a lot of time in preparing this LaTeX-document and optimizing the layout process in an excellent manner.

Finally, we would like to thank Dr. Werner A. Müller and Christian Rauscher from Springer-Verlag for their support in publishing this proceedings volume.

München,
January 2011

Bo Hu
Karl Morasch
Stefan Pickl
Markus Siegle

Contents

I.1 Forecasting, Data Mining and Machine Learning

Chair: Prof. Dr. Renato de Leone (Università di Camerino, Italy)

Original contributions addressing both theoretical aspects and practical applications in the areas of Forecasting, Data Mining and Machine Learning are solicited. This topic will focus on, but will not be limited to, methods and tools for analyzing, understanding and explaining large sets of data, searching for relationships in the data, forecasting and prediction.

Today increasing quantities of data are available to organizations and it is extremely important the ability of searching data in order to capture significant characteristics and extract meaningful patterns and strategic knowledge. A major challenge, and a great opportunity, is to effectively use all of this data to make valuable decisions. Machine Learning techniques provide useful tools for data classification and feature extraction and support for data forecasting.

Possible contributions may address theoretical contributions in this topic, new tools from mathematical programming and operations research, and novel modeling approaches as well as results of applications of these techniques and models to specific real problems.

A Bayesian Belief Network Model for Breast Cancer Diagnosis

S. Wongthanavasu

Abstract A statistical influence diagram, called Bayesian Belief Network (BBN), is investigated in modeling the medical breast cancer diagnosis. The proposed BBN is constructed under supervision by medical experts. Four types of datasets, namely, historic biodata, physical findings, indirect and direct mammographic findings are taken into consideration for modeling the BBN. Biodata are comprised of age, number of relatives having breast cancer, age at first live birth and age at menarche. Physical findings consist of pain, axilla, inflame and nipple discharge. Indirect mammographic data are breast composition. Direct mammographic findings are information obtained by mammogram image processing using the proposed cellular automata algorithms. A dataset is collected in real case of the breast cancer patients who come to get serviced at Srinakarind Hospital, Khon Kaen University, Thailand. A dataset of 500 cases is used throughout for model's performance evaluation.

In this regard, an 80 % of data is used for training the model, while the rest of 20 % is utilized for testing. The trained BBN model is tested on 100 patients consisting of 50, 25 and 25 for normal, benign and malignant patients, respectively. The proposed BBN provides the promising results reporting the 96.5 % of accuracy in the diagnosis. In addition, 5-fold and 10-fold cross-validation approach are implemented, the proposed BBN reports the promising results. It provides 96.2 and 97.4 percentages of accuracy, respectively.

1 Problem Overview

Mammography is an important tool in early detection of breast cancer. Unfortunately, many mammography findings cannot be classified easily as malignant or benign.

Wongthanavasu, S.
Department of Computer Science, Faculty of Science
Khon Kaen University, Khon Kaen, Thailand
e-mail: wongsar@kku.ac.th

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_1,

Successful diagnosis depends on the ability of a physician to detect mammographic abnormalities and to integrate clinical information such as risk factors and physical findings to determine the likelihood of breast cancer. In this regard, machine learning are capable to successfully assist physicians in detecting the early detection of breast cancer. Bayesian network is one of the promising candidates. There are a number of papers that reported applications of machine learning techniques in breast cancer diagnosis [1, 2, 9]. In this respect, some papers to date investigates breast cancer diagnosis using Bayesian networks [4, 5, 6]. Bayesian Belief Network (BBN) is a promising machine learning technique superior to well-known techniques such as support vector machines and neural networks in several reasons. Firstly, BBN is capable of predicting and diagnosing disease by using incomplete input data while the two previously stated methods can not. Secondly, support vector machines and neural networks can not perform diagnosis besides prediction. That is they do not accept incomplete data for the test. Thirdly, there are many medical diagnosis models in literature report to date that BBN is superior in medical diagnosis problems. In this regard, there are several papers reports the investigation of breast cancer using the BBN. Consequently, the modeling is the key successful issue.

This paper proposes a breast cancer diagnosis model using Bayesian belief network. The major issues of the model are an integration of mammographic with clinical information.

2 Model Structure

2.1 Bayesian Belief Networks

A Bayesian network, also called a belief network or causal probabilistic network, is a graphical representation of probabilistic information: it is a directed, acyclic graph in which nodes represent random (stochastic) variables, and links between nodes represent direct probabilistic influences between the variables [5]. In this formalism, propositions are given numerical probability values signifying the degree of belief accorded them, and the values are combined and manipulated according to the rules of probability theory. Each node represents a variable and has two or more possible states. For example, the variable "Breast Cancer" has two states: "present" and "absent". Each state is associated with a probability value; for each node, these probability values sum to 1.

For implementing the proposed BBN, nodes and their states are enumerated in Table 1. Four types of datasets, namely, patient history, physical findings, indirect mammographic findings, and direct mammographic findings were used in the BBNs modeling. In this regard, the associated network was shown in Fig. 1.

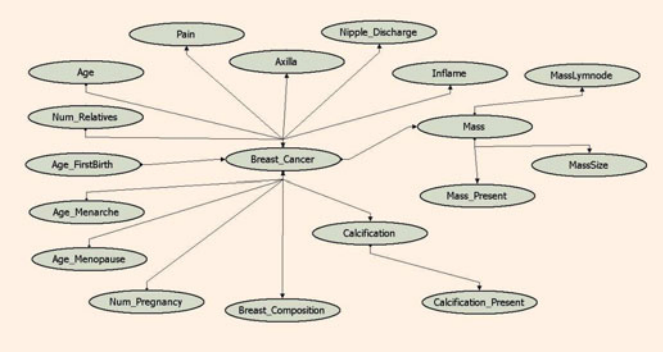


Fig. 1 Proposed BBN Model.

Table 1 Definition of BBN model's nodes (variables) and their states

Category	Node (variables)	States
Diagnosis	Age (years)	<20, 20-30, 31-40, 41-50, 51-60, >60
Patient History	Breast Cancer	present, absent
	Age at Menarche (years)	<12, 12-13, >13
	Age at First Live Birth (years)	<20, 20-24, 25-29, >29
	Number of First-Degree Relatives with Breast Cancer	0, 1, 2
	Age at First Live Birth (years)	<20, 20-24, 25-29, >29
	Age at Menopause (years)	<40, 40-44, 45-40, >49
	Number of Pregnancy	<2, 2-3, >3
Physical Findings	Pain	present, absent
	Nipple Discharge	present, absent
	Axilla	present, absent
	Inflame	present, absent
Indirect Mammographic Findings	Breast Composition	present, absent
Direct Mammographic Findings	Mass	malignant, benign, none
	Mass Present	yes, no
	Mass Lympnode	present, absent
	Mass Size (cm.)	1,2,3,4
	Calcification	malignant, benign, none
	Calcification Present	yes, no

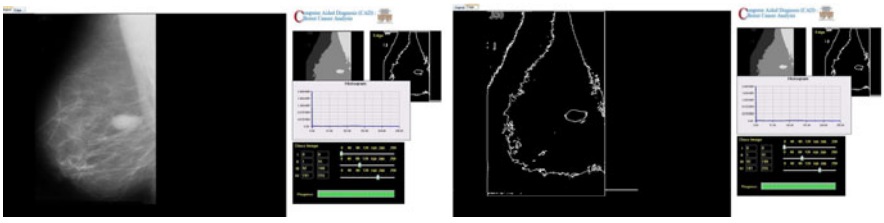
2.2 Direct Mammographic Findings

Mammographic image processing was carried out using cellular automata model (CA) [7, 8, 9, 10] to determine mass, calcification, and their features. Fig. 2 given below shows a process of mass detection in Computer Aided Diagnosis (CAD) software due to the research project.

2.3 Data Acquisition and Inference Software

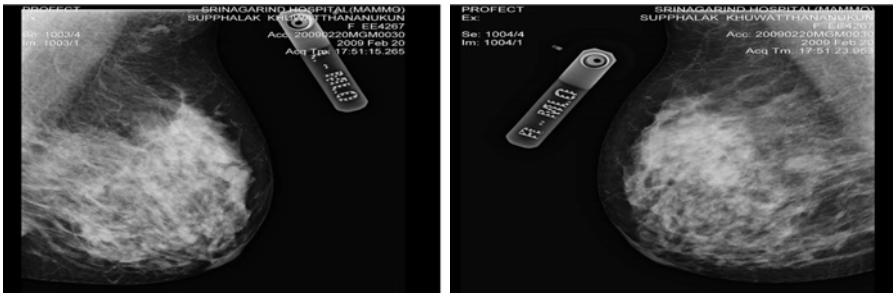
The proposed BBN's knowledge base was constructed from medical literature, health statistics reports and 500 patient's records collected at Srinakarind Hospital, Khon Kaen University. Reasoning in MSBNx inference software [3] uses conditional independence. It means that each variable (node) is independent with other variables which are not its successor given direct predecessors. This condition makes statistical reasoning possible by reducing combinatorial explosion of joint probability distribution. The patient dataset is calculated for conditional probabilities using conditional independence for each variable in the proposed BBN in Fig. 2. Fig. 2 shows mammograms of a benign patient.

Microsoft Bayesian Networks (MSBNx) [3] was utilized as a software tool to model the breast cancer diagnosis problem. MSBNx is a component-based Windows application for creating, assessing, and evaluating Bayesian Networks developed by Microsoft company. It is firstly developed as tools for assisting helpdesk in diagnosis of user's Microsoft office problems. MSBNx can deal with categorical data only by using conditional independence in reasoning. In implementation the proposed BBN model using MSBNx, 100 patients comprised of 50 normal, 25 benign, and 25 malignant were tested on the proposed model. In this regard, it provides the promising results by reporting the percentage of 96.5 of accuracy in the diagnoses.



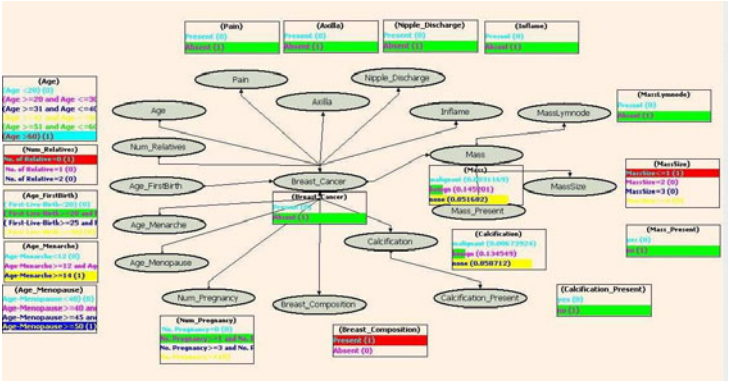
(a) Before processing (b) After processing

Fig. 2 Mammogram Image Processing Using Cellular Automata Model

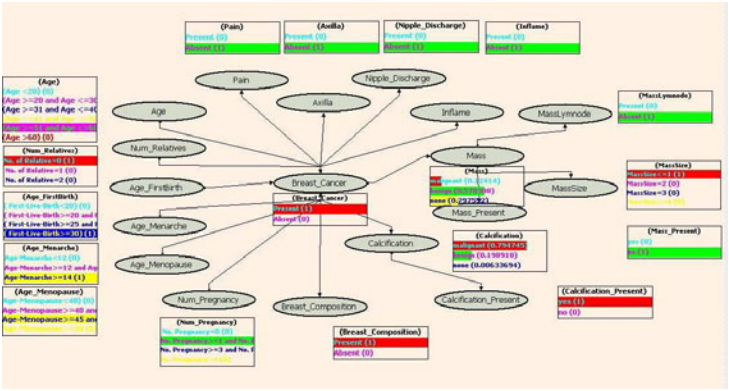


(a) Left (b) Right

Fig. 3 Mammograms for a benign patient



(a) Benign patient



(b) Malignant patient

Fig. 4 An example of BBN's results for diagnosing the test patients: a) benign, and b) malignant

Figure 1 shows the results obtained from the system as implemented in three cases of patients.

In addition, 5-fold and 10-fold cross-validation approach are implemented on such a dataset of 500 cases. On average, the proposed model reports 96.2 and 97.4 percentage of accuracy for 5-fold and 10-fold cross-validation, respectively.

3 Conclusions and Discussions

Bayesian networks represent a promising technique for clinical decision support and provide a number of powerful capabilities for representing uncertain knowledge. They provide a flexible representation that allows one to specify dependence and independence of variables in a natural way through the network topology.

Because Bayesian networks represent uncertainty using standard probability, one can collect the necessary data for the domain model by drawing directly on published statistical studies or elicited from the experts. The proposed BBN model is undergoing preclinical testing to compare its performance to that of radiologists with varying levels of mammographic expertise. In addition, we are considering the addition of variables to improve the model performance, such as demographic features such as race and geographic location, and patient-history features such as diet, body habitus, history of hormone therapy, and previous cancers. We hope to report the investigation soon.

Acknowledgements We thank The Thailand Research Fund (TRF) and The Office of Higher Education Commission for the financial support of the research through RMU5080010.

References

1. Cheng HD, Shan J, Ju W, Guo Y, Zhang L (2010) Automated Breast Cancer Detection and Classification Using Ultrasound Images: A Survey. In: Pattern Recognition, vol 43, pp 299–317
2. Huang CL, Liao HC, Chen MC (2008) Predication Model Building and Feature Selection with Support Vector Machines in Breast Cancer Diagnosis. Expert Systems with Applications. vol 34, pp 578–587
3. Locked J (1999) Microsoft Bayesian Network: Basics of Knowledge Engineering, Microsoft Technical Report
4. Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN (2008) A Bayesian Derived Network of Breast Pathology Co-occurrence. In: Journal of Biomedical Informatics, vol 41, pp 242–250
5. Nicandro CR, Héctor GAM, Humberto CC, Luis ANF, Rocío EBM (2007) Diagnosis of breast cancer using Bayesian networks: A case study. Computers in Biology and Medicine. vol 37, pp 1553–1564
6. Wang XH, Zheng B, Good WF, King JL, Chang YH (1999) Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. In: International Journal of Medical Informatics, vol 54(2), pp 115–126
7. Wongthanavasu S, Lursinap C (2004) A 3-D CA-based Edge Operator for 3-D Images. In: Proceedings of the 11th IEEE Int. Conf. Image Processing (IEEEICIP 2004), pp 235–238
8. Wongthanavasu S, Sadananda R (2000) Pixel-level Edge Detection Using a Cellular Automata-Based Model. In: Proceedings of Int. Conf. Advances in Intelligent Systems: Theory and Applications (59), IOS Press, The Netherlands, pp 343–351
9. Wongthanavasu S, Tangvoraphonkchai V (2007) Cellular Automata-based Algorithm and Its Application in Medical Images Processing. In: Proceedings of the 14th IEEE Int. Conf. Image Processing (IEEE-ICIP 2007), pp III–41–44
10. Wongthanavasu S, Tanvoraphonkchai V (2008) Cellular Automata-based Identification of The Pectoral Muscle in Mammograms. In: Proceedings of the 3rd Int. Conf. Biomedical Engineering (ISBME 2008), pp 294–298

A Method for Robust Index Tracking

Denis Karlow and Peter Rossbach

1 Introduction

In today's Portfolio Management many strategies are based on the investment into indices. This is a consequence of various empirical studies that show that the allocation over asset classes, countries etc. provides a greater performance contribution than the selection of single assets. For every portfolio containing indices as components the problem is that an index cannot be purchased directly. So it has to be rebuilt. This is called index tracking. The goal is to approximate the risk and return profile of an index. There exist different approaches to track indices [6]. One widely used approach is sampling, where the index is reproduced by a so-called tracking portfolio with a smaller number of assets, mostly a subset of its components. When applying sampling one has to solve two sub problems: selecting the assets of the tracking portfolio and determining their fractions. The quality of tracking is measured by the tracking error, which expresses the deviation between the returns of the tracking portfolio and the returns of the index. The optimal tracking portfolio is the portfolio with the smallest tracking error among all possible tracking portfolios. This is usually calculated with past data, assuming that the tracking quality will remain in the future investment period.

In the remainder of the paper, we show that the common approaches of sampling have weaknesses to generate tracking portfolios with a stable quality even in the estimation and investment period (section 2).

D. Karlow
Frankfurt School of Finance & Management, Sonnemannstrasse 9–11, DE-60314 Frankfurt am Main, Germany. e-mail: d.karlow@fs.de

P. Rossbach
Frankfurt School of Finance & Management, Sonnemannstrasse 9–11, DE-60314 Frankfurt am Main, Germany. e-mail: p.rossbach@fs.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_2, © Springer-Verlag Berlin Heidelberg, 2011

We will then introduce a new approach based on the technique of support vector regression to overcome this weakness (section 3). Finally, empirical results comparing our approach and the traditional approach are presented (section 4).

2 Weaknesses of Traditional Sampling Methods

To calculate the tracking error, different measures are used in literature. [5] show that the mean absolute deviation (MAD) has some advantages compared to the alternatives, like mean square error. It is defined by:

$$MAD = \frac{1}{T}\sum_{t=1}^{T}|R_{P,t} - R_{I,t}| \tag{1}$$

where $R_{P,t}$ and $R_{I,t}$ are the returns of a tracking portfolio and the index accordingly at time t, T is the numbers of time periods.

Applying sampling means to first choose the components of the tracking portfolio and second calculate the fractions of these components. The latter is usually done by calculating the weights that provide the minimal tracking error using an optimization algorithm. Choosing the components results in a combinatorial problem which can be solved using simple heuristics, like market capitalization, or higher sophisticated methods like genetic algorithms. In both cases, the assumption is that a tracking portfolio with a high tracking quality in the estimation period has also a high quality in the investment period.

Our empirical findings do not confirm this assumption. In a first experiment, the goal was to track the DAX30 index with different numbers N of stocks. To choose these stocks, we used two methods. One method always uses the N stocks with the highest capitalization within the DAX and alternatively a Genetic Algorithm (GA) to select the N stocks according to their quality of tracking in the estimation period. In both variants the weights for a given portfolio structure were calculated by minimizing the MAD. We applied this experiment for many different points of time and $N \in [10, 15, 20]$. In every case, the GA was the best in the estimation period, but in the investment period it was worse in nearly half of the cases.

To analyze the stability of the tracking quality, in a second experiment we calculated the optimal weights of all possible tracking portfolios at a given point of time, ranked them according to their tracking quality in the estimation and in the investment period and finally allocated the portfolios into percentiles for both periods according to the ranks. The results show, that the majority of portfolios do not remain in the same percentile. Hence, a good tracking quality in the estimation period does not guarantee also a good quality in the investment period. Thus, there is a need for a method to calculate the fractions of the components of a tracking portfolio that provides a better generalization.

3 Index Tracking Using Support Vector Regression

We propose a method that provides such a generalization ability. It is based on the support vector regression (SVR) method which is a special type of the support vector machines developed by [9]. The SVR uses a variant of the mean absolute deviation measure. This so-called ε-insensitive loss function does not penalize acceptable deviations defined by a parameter ε:

$$|R_{P,t} - R_{I,t}|_\varepsilon = max\{0, |R_{P,t} - R_{I,t}| - \varepsilon\} = \xi_t^*, t = 1, .., T. \tag{2}$$

As a consequence, there exist a region with a loss of zero between $+\varepsilon$ and $-\varepsilon$ and a region with a loss outside, expressed in the error slopes ξ_t, ξ_t^*. The regression based on this loss function has the following form:

$$\begin{aligned}
\underset{w}{\text{minimize}} \quad & \frac{C}{T}\sum_{t=1}^{T}(\xi_t + \xi_t^*) + \frac{1}{2}\|w\|_2^2 \\
\text{subject to} \quad & R_{I,t} - \sum_{i=1}^{N} w_i r_{i,t} \leq \varepsilon + \xi_t \\
& \sum_{i=1}^{N} w_{i,t} r_{i,t} - R_{I,t} \leq \varepsilon + \xi_t^* \\
& \sum_{i=1}^{N} w_{i,t} = 1, w_{i,t} \geq 0, i = 1, .., N \\
& \xi_t, \xi_t^* \geq 0 \\
& t = 1, .., T.
\end{aligned} \tag{3}$$

where $r_{i,t}$: return of asset i at time t, $w_{i,t}$: weight of asset i at time t, $t = 1, ..., T$: time periods, $i = 1, ..., N$: number of shares, ε: insensitive region (preset), C: regularization parameter (preset).

The minimization of the error slopes leads to a good approximation of the index by the tracking portfolio. Varying the size of the insensitive region can help to achieve a good generalization. Reducing the size of the region increases the approximation error and may lead to overfitting. In contrast, increasing the size of the region reduces the approximation error and may lead to underfitting. Thus, an optimal trade-off between approximation and generalization expressed by the parameters C and ε must be found. This is an ill-posed problem, which could have many solutions. Therefore, the introduction of $\|w\|_2$ helps to find a unique solution.

In most tracking approaches the optimal tracking portfolio is calculated by finding the optimal weights. The further investment is then based on these weights. The problem is that this is not consistent with a "buy and hold" strategy, because relative price changes lead to transactions to keep the weights constant [1]. Thus, fixing the numbers instead of the weights seems to be appropriate. But weights are necessary for specific types of constraints.

Assuming, that the relations between the final value of the tracking portfolio and any of its values in time are nearly equal to the corresponding relations of the index [4], we calculate a correction factor which turns the fixed weights into variable weights and therefore into fixed numbers:

$$w_{i,t} = w_{i,T}\frac{V_{P,T}}{V_{P,t}}\frac{S_{i,t}}{S_{i,T}} = [\frac{V_{P,T}}{V_{P,t}} \approx \frac{V_{I,T}}{V_{I,t}}] = w_{i,T}\frac{V_{I,T}}{V_{I,t}}\frac{S_{i,t}}{S_{i,T}} \tag{4}$$

where $V_{I,t}$: value of the index at time t, $V_{P,t}$: value of the tracking portfolio at time t, $S_{i,t}$: price of asset i at time t. The final formulation of our index tracking model is:

$$\begin{aligned}
\underset{w}{\text{minimize}} \quad & \frac{C}{T}\sum_{t=1}^{T}(\xi_t + \xi_t^*) + \frac{1}{2}\|w\|_2^2 \\
\text{subject to} \quad & R_{I,t} - \sum_{i=1}^{N} w_{i,T}[\frac{V_{I,T}}{V_{I,t}}\frac{S_{i,t}}{S_{i,T}}]r_{i,t} \leq \varepsilon + \xi_t \\
& \sum_{i=1}^{N} w_{i,T}[\frac{V_{I,T}}{V_{I,t}}\frac{S_{i,t}}{S_{i,T}}]r_{i,t} - R_{I,t} \leq \varepsilon + \xi_t^* \\
& \sum_{i=1}^{N} w_{i,T} = 1, w_{i,T} \geq 0, i = 1,..,N \\
& \xi_t, \xi_t^* \geq 0 \\
& t = 1,..,T.
\end{aligned} \tag{5}$$

For the selection of the parameters C and ε we apply the structural risk minimization approach [9]. When applying the SVR, the data is divided into two subsets, a training set (estimation period) to estimate the model and a test set (investment period) to validate the model. The empirical error reflects the fitting quality of the model in the estimation period. The test error reflects the generalization ability of the model in the investment period. The choice of the parameters C and ε helps to reduce this test error. For a classification problem it is bounded by:

$$R(w) \leq R_{emp} + \Phi(T/h) \tag{6}$$

where $R(w)$: test error, R_{emp}: empirical error, $\Phi(T/h)$: confidence interval, T: sample size, h: complexity measure (i.e. degree of freedom).

Assuming, that there is an optimal complexity for a model, increasing the complexity would result in an overfitting and vice versa. The confidence interval $\Phi(T/h)$ is larger for complex models and smaller for simple models. The goal is to select a model with a complexity that minimizes the sum of the empirical error and the confidence interval. This can be achieved by controlling the model complexity h, which depends on the parameters C and ε. To estimate h for different C and ε we use a procedure described in [8] with extensions from [7]. This procedure is defined for a classification problem. [3] have shown that a support vector regression problem has an equivalent support vector machine classification problem.

Applying this, we measure the model complexity using the training data to find optimal values for C and ε to minimize the boundary for the test error (6) and therefore to find the optimal generalization ability. We utilize this tuning procedure for (5).

4 Empirical Results

To validate the approach empirically, we use the data from the OR Library of [2]. We apply our method to the DAX100 and the S&P500. The time series of every index is divided into two subsets of equal size, one for the estimation period and one for the investment period. We select the minimum and maximum size for the tracking portfolios (TP) and divide the resulting range into steps of 5 points for the DAX100 respectively 10 points for the S&P500. Finally, for every of the resulting portfolio size we randomly select the stocks and calculate the optimal tracking portfolios using the traditional MAD and our approach. The last step is repeated 100 times for every tracking portfolio size.

Both used times series have different characteristics. While the series of the DAX100 has an upward trend over the whole time period, the S&P500 has just an upward trend in the estimation period which turns to a downward trend with the beginning of the investment period. The results of tracking the DAX100 (see table 1) show that in the estimation period our approach is worse compared to the MAD. But in the investment period it slightly outperforms the MAD. Due to the same trend in the estimation and the investment period the difference is not large. Regarding the results of the S&P500 (see table 2) our approach is again worse in the estimation period. But in the investment period it clearly outperforms the MAD. To prove the significance of the results we compare the means of the tracking errors for both methods in the investment period using a t-test.

Table 1 Means and standard deviations of tracking errors for the DAX100

	Tracking Error for MAD				Tracking Error for SVR				
TP	Estimation		Investment		Estimation		Investment		p-
Size	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	value*
10	0.633	0.137	0.913	0.185	0.661	0.136	0.897	0.170	0.256
15	0.466	0.081	0.752	0.124	0.501	0.087	0.732	0.121	0.124
20	0.402	0.089	0.675	0.107	0.437	0.096	0.659	0.106	0.152
25	0.317	0.057	0.574	0.078	0.350	0.065	0.561	0.073	0.123
30	0.290	0.062	0.537	0.091	0.321	0.068	0.529	0.085	0.252
35	0.254	0.052	0.502	0.080	0.283	0.057	0.491	0.076	0.168
40	0.223	0.042	0.473	0.070	0.252	0.047	0.458	0.063	0.055
45	0.205	0.047	0.445	0.067	0.233	0.054	0.433	0.063	0.113
50	0.172	0.037	0.399	0.060	0.197	0.045	0.389	0.055	0.109

* t-test for investment period: $H_0 : \mu_{svr} \geq \mu_{mad}, H_1 : \mu_{svr} < \mu_{mad}$

The last columns in the tables show the resulting p-values, which can be interpreted as the probability that the SVR is worse or equal to the MAD.

Table 2 Means and standard deviations of tracking errors for the S&P500

	Tracking Error for MAD				Tracking Error for SVR				
TP	Estimation		Investment		Estimation		Investment		p-
Size	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	value*
20	0.732	0.095	1.484	0.299	0.773	0.101	1.442	0.285	0.161
30	0.583	0.072	1.247	0.212	0.637	0.075	1.228	0.226	0.273
40	0.511	0.057	1.119	0.188	0.569	0.067	1.096	0.219	0.220
50	0.446	0.052	1.021	0.144	0.505	0.059	1.002	0.144	0.177
60	0.412	0.050	0.954	0.133	0.472	0.054	0.922	0.137	0.050
70	0.380	0.046	0.912	0.133	0.437	0.053	0.882	0.152	0.072
80	0.342	0.039	0.875	0.133	0.391	0.048	0.829	0.104	0.004
90	0.319	0.033	0.830	0.116	0.366	0.041	0.785	0.129	0.005
100	0.305	0.037	0.798	0.119	0.348	0.052	0.753	0.121	0.005

* t-test for investment period: $H_0 : \mu_{svr} \geq \mu_{mad}, H_1 : \mu_{svr} < \mu_{mad}$

5 Conclusion

The results indicate two advantages of our SVR-based approach. First, the results approve that our approach has the better generalization ability compared to the MAD, meaning that problems resulting from overfitting can explicitly be handled. Second, our approach shows that it is more robust to changes in trend.

References

1. G. Bamberg and N. Wagner. Equity index replication with standard and robust regression estimators. *OR Spectrum*, 22(4): 525–543, 2000.
2. J.E. Beasley. OR-Library: Distributing test problems by electronic mail. *Journal of the Operational Research Society*, pages 1069–1072, 1990.
3. J. Bi and K.P. Bennett. A geometric approach to support vector regression. *Neurocomputing*, 55(1-2): 79–108, 2003.
4. K. Montfort, E. Visser, and L.F. van Draat. Index tracking by means of optimized sampling. *The Journal of Portfolio Management*, 34(2): 143–152, 2008.
5. M. Rudolf, H.J. Wolter, and H. Zimmermann. A linear model for tracking error minimization. *Journal of Banking & Finance*, 23(1): 85–103, 1999.
6. S.A. Schoenfeld. *Active index investing*. Wiley, 2004.
7. X. Shao, V. Cherkassky, and W. Li. Measuring the VC-dimension using optimized experimental design. *Neural computation*, 12(8): 1969–1986, 2000.
8. V. Vapnik, E. Levin, and Y.L. Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5): 851–876, 1994.
9. V. N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

Forecasting Complex Systems with Shared Layer Perceptrons

Hans-Jörg von Mettenheim, Michael H. Breitner

Abstract We present a recurrent neural network topology, the Shared Layer Perceptron, which allows robust forecasts of complex systems. This is achieved by several means. First, the forecasts are multivariate, i. e., all observables are forecasted at once. We avoid overfitting the network to a specific observable. The output at time step t, serves as input for the forecast at time step t+1. In this way, multi step forecasts are easily achieved. Second, training several networks allows us to get not only a point forecast, but a distribution of future realizations. Third, we acknowledge that the dynamic system we want to forecast is not isolated in the world. Rather, there may be a multitude of other variables not included in our analysis which may influence the dynamics. To accommodate this, the observable states are augmented by hidden states. The hidden states allow the system to develop its own internal dynamics and harden it against external shocks. Relatedly, the hidden states allow to build up a memory. Our example includes 25 financial time series, representing a market, i. e., stock indices, interest rates, currency rates, and commodities, all from different regions of the world. We use the Shared Layer Perceptron to produce forecasts up to 20 steps into the future and present three applications: transaction decision support with market timing, value at risk, and a simple trading strategy.

Hans-Jörg von Mettenheim, e-mail: mettenheim@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

Michael H. Breitner, e-mail: breitner@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_3,

1 Methodology

Today's financial markets are complex. This is mostly due to markets not moving independently from one another. Instead, financial assets are correlated, see for example [6]. Additionally, it is generally not possible to assign clear causal relationships. A very simple example: we often assume that rising interest rates make a currency more attractive and cause it to appreciate. However, rising interest rates can also signify that inflation is high and this is more likely to cause the currency to depreciate. Also, the other way around, an appreciating currency can trigger rising *or* falling interest rates depending on the situation. And this is just a possible 1:1 relationship. When considering still other asset classes like stock indices or commodities relationships become more and more difficult to identify.

For this reason we use a novel recurrent neural network architecture, due to [8, 9]. Figure 1 shows the network: at every time step t we have a state vector s^t. The upper part of s^t contains the actual observables. These are symbolized by bold circles in the figure. The lower part of s^t contains hidden states. These states serve as memory, allow the system to develop its own internal dynamics and hardens against external shocks. The transition from time t to $t+1$ occurs via the transition equation

$$s^{t+1} = \tanh(W \cdot s^t) \tag{1}$$

where W is a sparse matrix of weights. These weights are optimized during the training process. Note, that tanh is applied component-wise to the resulting vector.

Using this architecture has several benefits: it allows to forecast several time series, that is, all the observables, *at once*. There is no dedicated input or output. It offers the possibility to achieve multi step forecasts by repeatedly applying equation 1. By training several hundred networks we also get a distribution of possible outcomes. This allows to gauge the risk of the resulting forecast, see also the application in figure 2. This architecture also leads to robust models. Trying to match several time series reduces the risk of overfitting to a specific time series.

To model financial markets we have to select observables. As there is no commonly agreed on economic theory of how the markets work we select observables which we deem representative, see also [4]. These include stock indices, major currency pairs, benchmark interest rates for three months and ten years, and a selection of commodities, see [7]. Finally, the interest rates are used to create a crude yield curve. In total our model includes 25 observables.

Our dataset contains ten years of daily observation from Datastream. It spans the time from July, 1st, 1999 to June, 30th, 2009. For training, we first transform the data into returns or first differences in the case of yield curves. For forecasting we use a rolling window model: we train and validate our model on 440 days of past data. Then, we forecast on the following 110 days. Finally we move forward one month, that is 22 trading days, and retrain. We train an ensemble of 300 networks. Our weight matrix uses a sparsity of 12.5 %.

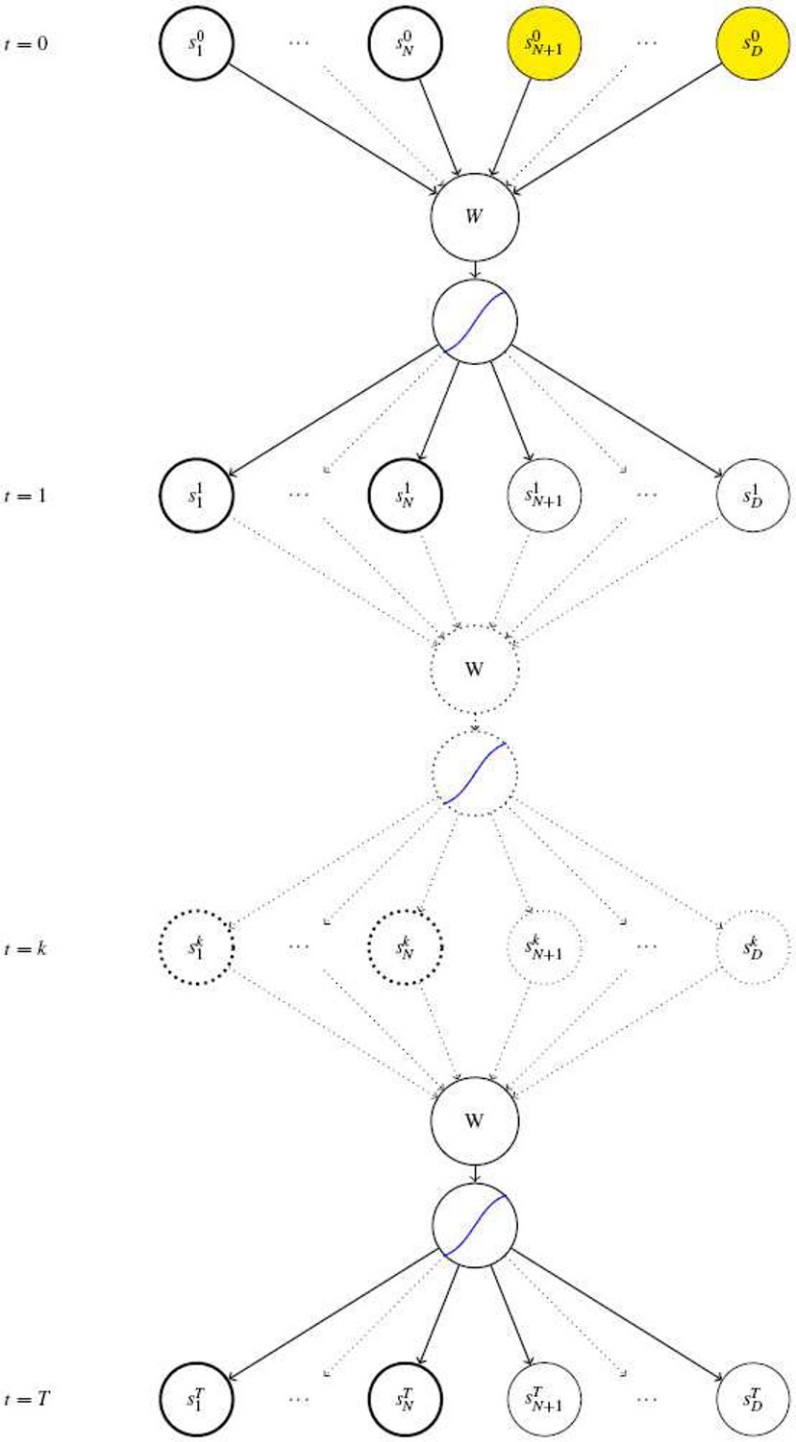


Fig. 1 This recurrent neural network architecture models several time series at once.

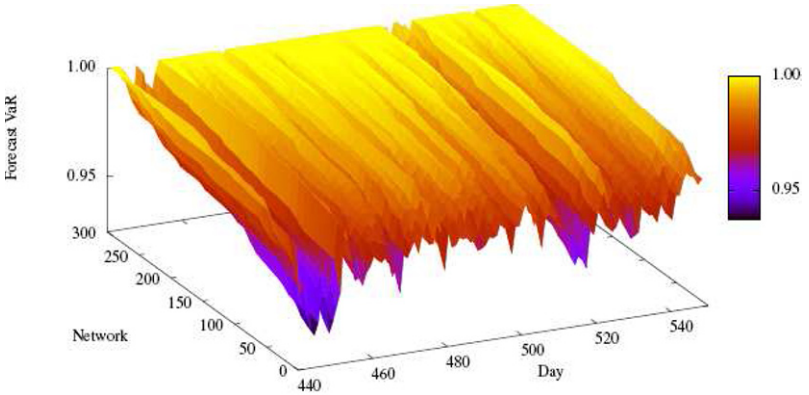


Fig. 2 Example of an application: forecasting Value at Risk (VaR) for the GBP|USD exchange rate. The figure shows the distribution of an expert topology of 300 networks. The forecast starts at day 441 because the first 440 days are used for training.

2 Applications

Our applications belong to the class of decision support systems. For a description of neural networks in this context see also [1]. The first application deals with Value at Risk, see [5]. Value at Risk is a very important measure among financial risk managers. It answers the question: how much is my portfolio likely to loose at most within the next N days with a probability of X percent. Using this terminology our application deals with a 10 days, 95 percent Value at Risk. Getting Value at Risk right is important for two reasons: on the one hand, violating Value at Risk shows that risk is not managed correctly and this will lead to sanctions by the regulation authorities. On the other hand, estimating too conservative a Value at Risk leads to capital being bound which one could otherwise put to better use elsewhere.

Figure 2 exemplarily shows how we can calculate Value at Risk from an ensemble of networks. Here we use the exchange rate of Great British Pound to US Dollar (GBP|USD). At every time step we get a forecast of portfolio returns over the next ten days. Because we use an ensemble of 300 networks we also get 300 returns. By ordering the returns from best to worst we get a distribution from which we can read the desired Value at Risk. For example, we obtain the 95 percent Value at Risk by cutting the distribution at network 15, because $15 = 300 \cdot (1 - 0.95)$.

When calculating Value at Risk for all assets in our portfolio we notice that the violations are within the permitted range. Tracking is far better than popular benchmarks like, e. g., historical simulation. We present a comparison to statistical methods in [2].

Our second application concerns transaction decision support. The situation is the following: on an ongoing basis an asset has to be purchased. For example the task could involve buying a certain amount of a currency or a commodity every month.

Corporations with regular foreign cash flows or depending on a specific commodity are in this situation. The challenge is to buy at the lowest possible price in a given month. This task is also called market timing.

The presented network allows for multi step forecasts. To get a candidate for the time of the lowest price we compute trajectories of the given asset for every network of the ensemble. By averaging the results from all ensembles we get the desired price. We compute the realized potential of a given transaction. Realized potential measures how close we are to the lowest entry point which would have been available with perfect hindsight. A perfect match has a realized potential of 1.0. If we buy at the highest point in a given month realized potential is 0.0. As a benchmark we take a fixed day strategy where we buy on the first, second, ..., last trading day of a month and compute the excess realized potential. Considering a time span of 110 days our model always manages to beat all fixed day strategies regardless of the asset. When we let the model run on a time span of eight years without retraining we still beat the benchmarks most of the time. Note, that this is purely an exercise to see how robust the model is. Normally, a user would retrain the model.

Our third application belongs to investment decision support. It is a simple trading application whereby we are only interested in correctly forecasting the sign of the return for all assets, see also [3]. We choose to include this task because it is a common test for neural networks. The performance is easy to measure, for example as annualized return or as risk adjusted return. As is common we compare neural network performance against a naive strategy and a moving average strategy, see also [4].

As it turns out the neural network forecast beats the benchmarks or comes as a close second for all assets and all time spans. Like in the previous application the results are not always best, but they are robust. On the other hand the benchmarks sometimes produce stellar performance for some assets but totally fail for others. While no strategy manages to offer satisfying performance on the long time span of eight years the neural networks at least manage to break even without retraining.

Possible variations of the standard trading strategy include setting a threshold for the absolute value of the forecast. Forecasts below the threshold are either disregarded or used to go flat. The presented architecture allows a second variation: as we are interested in not trading too often to avoid transaction cost we can look not only a tomorrow's forecast but also at the day after tomorrow. We only trade if both forecasts have the same sign. Both variations do not manage to increase raw annualized returns but they often increase risk adjusted returns.

To summarize: in all applications and for all assets our networks generally manage to beat the benchmark. Although the performance is not always best it is robust in the sense that there are no catastrophic failures. The model works well for all asset classes. To conserve space we have opted to describe our results only qualitatively. Numerical values are available from the authors upon request.

3 Conclusions and Outlook

We show that the presented model is able to produce robust forecasts for complex systems like financial markets. Three applications exercise all aspects of the network architecture: forecasting several time series at once, multi step forecasts and probability distributions. We argue that a model that reliably forecasts not only one time series but several is more trustworthy than a model fitted to a single time series. We consider this network architecture a useful addition to the forecaster's toolbox.

The network architecture can be enhanced by several features: it is, for example, easily possible to account for missing data. Missing data is simply replaced by the network's own forecast. Considering the transition equation as operator it is also possible to accommodate non regular time grids. Additionally, models with different kind of input data should prove interesting: it is possible to add time information that is known in advance, like day of the week or month of the year. Also, the analysis of intraday data for short term decision support should lead to new insights.

References

1. Michael H. Breitner, Frank Köller, Simon König, and Hans-Jörg von Mettenheim. Intelligent decision support systems and neurosimulators: A promising alliance for financial services providers. In H. Österle, J. Schelp, and R. Winter, editors, *Proceedings of the European Conference on Information Systems (ECIS) 2007, St. Gallen*, pages 478–489, 2007.
2. Michael H. Breitner, Corinna Luedtke, Hans-Jörg von Mettenheim, Daniel Rösch, Philipp Sibbertsen, and Grigoriy Tymchenko. Modeling portfolio value at risk with statistical and neural network approaches. In Christian Dunis, Michael Dempster, Michael H. Breitner, Daniel Rösch, and Hans-Jörg von Mettenheim, editors, *Proceedings of the 17th International Conference on Forecasting Financial Markets: Advances for Exchange Rates, Interest Rates and Asset Management, Hannover, 26–28 May 2010*, 2010.
3. Christian Dunis, Jason Laws, and Patrick Naïm. Applied quantitative methods for trading and investment. Wiley, Southern Gate, Chichester, 2003.
4. Christian L. Dunis, Jason Laws, and Georgios Sermpinis. Higher order and recurrent neural architectures for trading the EUR/USD exchange rate. *Quantitative Finance*, 10, 2010.
5. John C. Hull. *Options, Futures and Other Derivatives*. Prentice Hall, 6th edition, 2006.
6. Ashraf Laïdi. *Currency Trading and Intermarket Analysis*. Wiley, New Jersey, 2009.
7. Hans-Jörg von Mettenheim and Michael H. Breitner. Robust forecasts with shared layer perceptrons. In Christian Dunis, Michael Dempster, Michael H. Breitner, Daniel Rösch, and Hans-Jörg von Mettenheim, editors, *Proceedings of the 17th International Conference on Forecasting Financial Markets: Advances for Exchange Rates, Interest Rates and Asset Management, Hannover, 26–28 May 2010*, 2010.
8. Hans Georg Zimmermann. Forecasting the Dow Jones with historical consistent neural networks. In Christian Dunis, Michael Dempster, and Virginie Terraza, editors, *Proceedings of the 16th International Conference on Forecasting Financial Markets: Advances for Exchange Rates, Interest Rates and Asset Management, Luxembourg, 27–29 May 2009*, 2009.
9. Hans Georg Zimmermann. Advanced forecasting with neural networks. In Christian Dunis, Michael Dempster, Michael H. Breitner, Daniel Rösch, and Hans-Jörg von Mettenheim, editors, *Proceedings of the 17th International Conference on Forecasting Financial Markets: Advances for Exchange Rates, Interest Rates and Asset Management, Hannover, 26–28 May 2010*, 2010.

Analysis of SNP-Complex Disease Association by a Novel Feature Selection Method

G. Üstünkar, S. Özöğür-Akyüz, G.-W. Weber and Y. Aydın Son

Abstract Selecting a subset of SNPs (Single Nucleotide Polymorphism pronounced snip) that is informative and small enough to conduct association studies and reduce the experimental and analysis overhead has become an important step toward effective disease-gene association studies. In this study, we developed a novel methods for selecting Informative SNP subsets for greater association with complex disease by making use of methods of machine learning. We constructed an integrated system that makes use of major public databases to prioritize SNPs according to their functional effects and finds SNPs that are closely associated with genes which are proven to be associated with a particular complex disease. This helped us gain insights for understanding the complex web of SNPs and gene interactions and integrate as much as possible of the molecular level data relevant to the mechanisms that link genetic variation and disease. We tested the validity and accuracy of developed model by applying it to real life case control data set and got promising results. We hope that results of this study will support timely diagnosis, personalized treatments, and targeted drug design, through facilitating reliable identification of SNPs that are involved in the etiology of complex diseases.

G. Üstünkar
Middle East Technical University, Informatics Institute,Ankara, Turkey, e-mail: e145307@metu.edu.tr

S. Özöğür-Akyüz
Bahçeşehir University, Department of Mathematics and Computer Science, Beşiktaş, Istanbul, Turkey e-mail: sureyya.akyuz@bahcesehir.edu.tr

G.-W. Weber
Middle East Techncial University, Institute of Applied Mathematics, Ankara, Turkey, e-mail: gweber@metu.edu.tr

Y. Aydın Son
Middle East Techncial University, Informatics Institute, Ankara, Turkey, e-mail: yesim@metu.edu.tr

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_4,

1 Introduction

SNPs are markers[1] with high potential for population genetic studies and for spotting genetic variations responsible for complex diseases. However, the huge number of SNPs makes it neither practical nor feasible to obtain and analyze the information of all the SNPs on the human genome. Thus, selecting a subset of SNPs that is informative and small enough to conduct association studies and reduce the experimental and analysis overhead has become an important step toward effective disease-gene association studies.

There has been an extensive research on selecting the optimal set of SNPs. However, the efficiency of searching for an optimal set of SNPs has not been as successful as expected in theory. It is computationally impossible to perform an exhaustive search of association for four or more interacting SNPs when analyzing data sets for several hundred individuals and several hundred thousands of SNPs [5]. The problem of SNP selection has been proven to be NP-hard in general [1], and current selection methods possess certain restrictions and require use of heuristics for reducing the complexity of the problem.

2 Literature Review

The aim of the Informative SNP Selection approach is to find a subset of SNPs with minimal cardinality, whose allele information can explain the rest of SNPs in the candidate region under study to the greatest detail. This way only those SNPs in the SNP subset (called Informative SNPs) can be used in the association studies without any loss of information. One can classify the proposed approaches into three categories according to how they try to measure the allele information: (a) Haplotype Diversity based approaches, (b) Pairwise Association based approaches, (c) Predicting Tagged SNPs.

Haplotype Diversity based approaches are inspired by the fact that DNA can be partitioned into discrete blocks such that within each block high Linkage Disequilibrium [2] (LD) is observed and between blocks low LD is observed [4, 9]. As a result of this feature, number of distinct haplotypes consisting of the SNPs is very small across a population. Hence, one would try to find the subset of SNPs, which are responsible for the "limited haplotype diversity" in order to find the representative SNP set. To cope with this problem, efficient heuristics have been proposed using Dynamic Programming [11], Principal Component Analysis [7] and Greedy Algorithm [11].

Pairwise Association based approaches are based on the principle that all the SNPs in the target locus are highly associated with at least one of the SNPs in

[1] DNA sequence with a known location on a chromosome and associated with a particular gene or trait.

[2] Non-random association of alleles at two or more loci, not necessarily in the same chromosome.

the selected SNP subset due to LD. This way, although a SNP that can be used to predict a disease causing variant may not be selected as a SNP, the association may be indirectly assumed from the selected SNP that is highly associated with it. The common solution approach for these methods is to cluster the SNPs into different subsets and choose a representative SNP (or SNPs) from each cluster [2, 3].

Predicting Tagged SNPs is motivated by the idea of reconstructing the haplotype data from an initial set of selected SNPs in order to minimize the error of prediction for unselected SNPs. [6] proposed a dynamic programming approach that fixed the number of tag SNPs for each tagged SNP to 2. [8] improves over the current predicting based method by allowing multi-allelic prediction (instead of bi-allelic) and not restricting the number of tag SNPs. Our selection method falls in this category.

3 Proposed Methodology

3.1 Filtering Biologically Relevant SNPs

Even with classification schemes that are optimized for large scale data it takes too much time to perform detailed analysis of a typical Genome Wide Association Study (GWAS) data. Therefore, there is a need to extract biologically relevant data as an initial pass. This would be achieved by incorporating information from biological databases so that biologically relevant SNPs are given higher priority. SPOT[3] recently introduced the genomic information network (GIN) method [10] for systematically implementing this kind of strategy. As a first step, we used SPOT for extracting biologically relevant SNPs from the whole SNP set after performing Case Control Association study to find p-values.

3.2 A Novel Informative SNP Selection Method

A formal definition of the problem can be stated as follows: Let $S = \{SNP_1, ..., SNP_n\}$ be a set of n SNPs in a candidate region and $G = \{g_1, ..., g_m\}$ be a data set of m genotypes, where each genotype g_i consists of the consecutive allele information of the n SNPs: $SNP_1, ..., SNP_n$ Suppose that the maximum number of SNPs that can be selected is k (which can be alternatively be a variable for the problem), and a function $f(R|G,P)$ evaluates how well the allele information of SNPs in subset $R \subset S$ retains the allele information of all SNPs in S based on the genotype data G and classification performance of selected set R on disease phenotype are. We set our goal as to find a minimum size set of representative SNPs and a prediction algorithm, such that the prediction error is minimized. Then our objective function becomes:

[3] http://spot.cgsmd.isi.edu/submit.php

$$\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{\tau_i}) + NaiveBayes(G_R, P),$$

where G_R denotes genotype data related with representative SNP set R, G_{τ_i} denotes genotype data related with $SNP_i \in S \setminus R$ and $NaiveBayes(F,L) = argmax_L P(L = l) \prod_{i=1}^{n} p)F_i = f_i | L = l)$ denotes a Naive Bayes classifier where F is the feature set (SNP set in our context) and L is the label. We used Simulated Annealing (SA) , which is a local search algorithm, to solve our problem. We create a random binary coding of size n as an initial solution and test the accuracy of the solution using Naive Bayes by calculating the mean classification error for $(n-k)$ supervised learning iterations, where k is the number of selected SNPs in a particular iteration. We run the algorithm for a certain amount of steps (user defined). We use a tradeoff between accuracy and the number of SNPs in the representative SNP set by introducing $E(s)$ as an evaluation function denoted by:

$$E(s) = w \left(\sum_{i=1}^{n-k} NaiveBayes(G_R, G_{\tau_i}) + NaiveBayes(G_R, P) \right) + (1-w)k.$$

Here, w denotes weight that specifies tradeoff. The smaller the w, the less SNPs will be chosen to represent overall SNP set S.

4 Experiments

Data at hand is whole genome association data for the North American Rheumatoid Arthritis Consortium (NARAC) including cases $(N = 868)$ and controls $(N = 1194)$ after removing duplicated and contaminated samples. It consists of 545080 SNP-genotype fields from the Illumina 550K chip. After initial quality control based filtering, size is reduced to 501.463 SNPs. We used SPOT for extracting biologically relevant SNPs. For our analysis we used multiple test adjusted p-values (False Discovery Rate) and used 0.25 as p-value threshold. We listed 9083 SNPs with biological relevance. In order to test the classification performance of the representative SNP set on phenotype variable, we first applied an initial split $(80\% - 20\%)$ on the filtered biologically and statistically relevant data and seperate training and testing set for our model. Therefore, we have randomly selected 1,650 patients for the training set and 588 patients for the test set for RA data. Following that, we ran our simulated annealing based representative SNP selection algorithm on the training set. As the algorithm is based on the idea of selecting the subset of SNPs, which best predicts the remaining SNPs for a *particular genomic region*; we ran the algorithm for each chromosome and merged the selected SNPs as the overall representative SNP set. The prediction performance (mean accuracy) of the selected SNPs (on unselected SNPs) for each chromosome is presented in Table 1 below:

Using representative SNP selection algorithm we managed to decrease the dimensions considerably. The number of SNPs is decreased from 9,083 to 1483 for

Table 1 Prediction Performance of Representative SNP Selection Algorithm: w = 0.7

Rheumatoid Arthritis

CN	I	S	PA	CN	I	S	PA
1	687	114	0.629	13	297	52	0.578
2	790	120	0.620	14	280	39	0.580
3	438	83	0.592	15	258	32	0.583
4	486	91	0.620	16	280	47	0.599
5	542	93	0.590	17	184	20	0.578
6	1.41	217	0.709	18	179	21	0.582
7	393	54	0.612	19	137	15	0.584
8	595	97	0.606	20	206	29	0.584
9	461	91	0.61	21	132	21	0.587
10	402	65	0.606	22	103	16	0.590
11	343	62	0.606	23	135	6	0.587
12	345	48	0.601	TOTAL	9083	1433	

[a]

[a] CN: Chromosome Number, I:Initial, S:Selected, PA:Prediction Accuracy

RA data set where as the average prediction accuracy (over not selected) of the selected SNP set for RA data is 0.601. To observe the classification performance of the selected set over the disease phenotype, we compared the performance against two filtering based attribute selection scheme from WEKA[4] tool set (Relief-F and Chi-Square) and randomly selected set of SNPs. In order to achieve that, we have selected the same set of SNPs for the test sets to that of training sets and applied a 10-fold Cross Validation (CV) run using Naive Bayes classifer as the supervised learning scheme.

Measures used for comparison purposes are Accuracy $((TP+TN)/(P+N))$, Recall $(TP/(TP+FN))$, Negative Predictive Value $(NPV = TN/(TN+FN))$, Precision $(TP/(TP+FP))$ and Specificity $(TN/(FP+TN))$ where TP denotes True Positive, TN denotes True Negative, FP denotes (False Positive) and FN denotes False Negative for a 2×2 confusion matrix. Results are presented in Table 2 below:

Table 2 10-Fold CV Results for RA Data

	w = 0.7, 1433 SNPs		
Measure	SA-SNP	Chi-Square	Relief-F
Accuracy	0,5728	0,5607	0,4587
Recall	0,0497	0,0000	1,0000
NPV	0,5689	0,5607	1,0000
Precision	0,6923	NA	0,4480
Specificity	0,9827	1,0000	0,0346

[4] (http://www.cs.waikato.ac.nz/ ml/weka)

Results reveal that our algorithm performs better against well known filtering based attribute selection schemes when prediction accuracy is favored against minimizing cardinality of SNP set.

5 Conclusion

During the course of this study, we introduced a novel Informative SNP Selection algorithm. We have developed an integrated system to help prioritizing SNPs according to biological relevance alongside with p-value of association. We have performed biological prioritization and SNP selection on real life data belonging to Rheumatoid Arthritis. We have performed a comparative study between two well known filtering based attribute selection methods with our method to test performance of our methodology. Our results revealed that we managed to reduce the dimension considerably without much information loss.

References

1. V. Bafna, B.V. Halldorsson, R. Schwartz, A.G. Clark, and S. Istrail. Haplotypes and informative SNP selection algorithms: don't block out information. In *Proc RECOMB'03*, pages 19–27, 2006.
2. M. C. Byng, J. C. Whittaker, A. P. Cuthbert, C. G. Mathew, and C. M. Lewis. SNP subset selection for genetic association studies. *Annals of Human Genetics*, 67: 543–556, 2003.
3. C.S. Carlson, M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet.*, 74: 106–120, 2004.
4. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. Highresolution haplotype structure in the human genome. *Nature Genetics*, 29(2): 229–232, 2001.
5. V. Dinu, H. Zhao, and P.L. Miller. Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. *J. of Biomedical Informatics*, 40(6): 750–760, 2007.
6. E. Halperin, G. Kimmel, and R. Shamir. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, 21(1): i195–i203, 2005.
7. B. Horne and N. J. Camp. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26(11): 11–21, 2004.
8. P. H. Lee and H. Shatkay. BNTagger: Improved Tagging SNP Selection using Bayesian Networks. *ISMB 2006 Supplement of Bioinformatics*, 22(14): e211–219, 2006.
9. N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, and D. P. McDonough. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294: 1719–1722, 2001.
10. S.F. Saccone, R. Bolze, P. Thomas, J. Quan, G. Mehta, E. Deelman, J.A. Tischfield JA, and J.P. Rice. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Research (Epub ahead of print)*, 2010.
11. K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. In *In Proceedings of the National Academy of Sciences (PNAS)*, volume 99, pages 7335–7339, 2002.

An Approach of Adaptive Network Based Fuzzy Inference System to Risk Classification in Life Insurance

Furkan Başer, Türkan E. Dalkılıç, Kamile Ş. Kula, Ayşen Apaydın

Abstract In this paper, we propose ANFIS based system modeling for classifying risks in life insurance. We differentiate policyholders on the basis of their cardiovascular risk characteristics and estimate risk loading ratio to obtain gross premiums paid by the insured. In this context, an algorithm which expresses the relation between the dependent and independent variables by more than one model is proposed to use. Estimated values are obtained by using this algorithm, based on ANFIS. In order to show the performance evaluation of the proposed method, the results are compared with the results obtained from the Least Square Method (LSM).

1 Introduction

The first recognition that fuzzy sets methodology could be applied to the problem of individual insurance underwriting was by [2]. In this paper, it is pointed out that the process of insurance underwriting, the process of selection and evaluation of risks to be insured, is subjective which may not be properly described by probability.

Furkan Başer, Department of Computer Applications Education, Faculty of Commerce and Tourism Education, Gazi University, Gölbaşı, 06830, Ankara, Turkey e-mail: fbaser@gazi.edu.tr, Tel.: +90 312 4851460 - 325; fax.: +90 312 4842316

Türkan E. Dalkılıç, Department of Statistics and Computer Sciences, Faculty of Arts and Sciences, Karadeniz Technical University, 61080, Trabzon, Turkey e-mail: tedalkilic@gmail.com

Kamile Ş. Kula, Department of Mathematics, Faculty of Arts and Sciences, Ahi Evran University,40200, Kırşehir, Turkey e-mail: kskula@ahievran.edu.tr

Ayşen Apaydın, Department of Statistics, Faculty of Science, Ankara University, Tandoğan, 06100, Ankara, Turkey e-mail: apaydin@science.ankara.edu.tr

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_5,

A basic form of the fuzzy expert system is used to analyze the underwriting practice of a life insurance company. [7] used a fuzzy expert system to provide a flexible definition of a preferred policyholder in life insurance. [10] used fuzzy expert system to model selection process in group health insurance.

In recent years; neural network, commonly used in analyzing complex problem, is emerged as an effective tool also in the area of parameter estimations. In addition to input variables having different distributions, observations for each variable may come from two or more classes and there may be some vagueness or fuzziness about the observations in data set belonging to these classes. There are too many methods in the literature for the estimation of unknown parameter. One of them is adaptive network based fuzzy inference system (ANFIS) [1]. ANFIS is a fuzzy interference tool implemented in the framework of adaptive network [6].

In this paper, we investigate an alternative method of classifying risks in life insurance, based on the concept of ANFIS based system modeling. We differentiate policyholders on the basis of their cardiovascular risk characteristics which are systolic blood pressure, cholesterol level, obesity and smoking behavior. By defining these risk factors as fuzzy variables, we estimate risk loading and the gross premium paid by the insured.

2 An Algorithm for Parameter Estimation Based ANFIS

There are several different types of fuzzy inference systems developed for function approximation. In this study, the Sugeno fuzzy inference system, which was proposed by [9], will be used. When the input vector X is $(x_1, x_2, ..., x_p)^T$ then the system output Y can be determined by the Sugeno inference system as:

$$R^L : If(x_1 \text{ is } F_1^L, and\ x_2 \text{ is } F_2^L, ..., and\ x_p \text{ is } F_p^L), then(Y = Y^L = c_0^L + c_1^L x_1 + ... + c_p^L x_p). \quad (1)$$

where F_j^L is fuzzy set associated with the input x_j in the L th rule and Y^L is output due to rule $R^L(L = 1, ..., m)$. The prediction of parameters with an adaptive network is based on the principle of the minimizing of error criterion. There are two significant steps in the process of prediction. First, we must determine a priori parameter set characterizing the class from which the data comes and then update these parameters within the process. The second step is to determine posteriori parameters belonging to the regression models [1, 3]. When the independent variables are coming from normal distribution, Gaussian membership function can be used. The algorithm related to the proposed method for determining the regression model in the case of independent variables coming from a normal distribution is defined as follows.

Step 1: Class numbers related to the data set associated with the independent variables are determined heuristically.

Step 2: The priori parameters are determined. Spreads and centers of the parameters are determined according to the intervals which the values of input variables

are included and according to the fuzzy class numbers of the variables. Then, the priori parameters are defined by:

$$v_i = \min(x_i) + \frac{\max(x_i) - \min(x_i)}{(l_i - 1)} * (i-1)\ i = 1, ..., p \tag{2}$$

Step 3: Weights, $\overline{W}^L$, are determined in order to form matrix B and to obtain the posteriori parameter set in the following step. The $\overline{W}^L$ weights are outputs of the nerves in the third layer of the adaptive network and they are counted based on a membership function related to the distribution family to which independent variable belongs. Nerve functions in the first layer of the adaptive network are defined by $f_{1,h} = \mu_{F_k}(x_i)\ h = 1, 2, ..., \sum_{i=1}^{p} l_i$ where $\mu_{F_h}(x_i)$ is called the membership function. Here, when the normal distribution function which has the parameter set of $\{v_h, \sigma_h\}$ is considered, these membership functions are defined as:

$$\mu_{F_h}(x_i) = exp\left[-\left(\frac{x_i - v_h}{\sigma_h}\right)^2\right] \tag{3}$$

The w^L weights are indicated as $W^L = \mu_{F_L}(x_i) \cdot \mu_{F_L}(x_j)$ and $\overline{w}^L$ weight is a normalization of the weight defined as w^L and they are counted with:

$$\overline{w}^L = \frac{w^L}{\sum_{L=1}^{m} w^L} \tag{4}$$

Step 4: The equation, $Z = (B^T B)^{-1} B^T Y$,is used to determine the posteriori parameter set, c_i^L. Here B is the data matrix which is weighted by membership degree. Y is dependent variable vector and Z is posterior parameter vector.

Step 5: By using a posteriori parameter set, c_i^L obtained in Step 4, the regression model indicated by $Y^L = c_0^L + c_1^l x_1 + c_2^L x_2 + ... + c_p^L x_p$ and the prediction values are obtained using

$$\hat{Y} = \sum_{L=1}^{m} \overline{w}^L Y^L \tag{5}$$

Step 6: The error related to model is obtained as:

$$\varepsilon = \frac{1}{n} \sum_{k=1}^{n} (y_k - \hat{y_k})^2 \tag{6}$$

If $\varepsilon < \phi$, then the posteriori parameters are handled as parameters of regression models, and the process is concluded. If $\varepsilon > \phi$, then, step 7 begins. Here ϕ , is a constant and determined by the decision maker.

Step 7: The priori parameters specified in Step 2 are updated with $v_i^{'} = v_i \pm t$. Here, t is the size of the step and a is a constant value.

Step 8: Predictions for each priori parameter obtained after revision and the error criteria related to these predictions are calculated. And then, the lowest error criterion is selected. The priori parameters giving the lowest error, and the prediction obtained via the models related to these parameters are taken as output.

3 Risk Evaluation by ANFIS

Traditionally, life insurance policyholders are classified by using classical mortality tables and generally according to limited number of risk characteristics, many other risk factors are ignored. Conventional clustering algorithms are organized in contemplation of the fact that objects are either in the set or not. However, the boundaries of each class are not always sharply defined. In these circumstances, fuzzy set methodology provides a convenient way for constructing a model that represents system more accurately since it allows integrating multiple fuzzy or non-fuzzy factors in the evaluation and classification of risks. Recent medical research states that cardiovascular diseases are the leading cause of illness and death. There are many risk factors associated with cardiovascular disease. The major risk factors, tobacco use, alcohol use, high blood pressure (hypertension), high cholesterol, obesity, physical inactivity, unhealthy diets, have a high prevalence across the world [8, 4].

In this paper, we differentiate policyholders on the basis of their systolic blood pressure, cholesterol level, obesity and the average cigarette consumption per day which are handled as fuzzy variables. Within this framework, systolic blood pressure is measured in mm of mercury (mm Hg) and level of cholesterol in mg per deciliter blood (mg/dl). Obesity is measured according to the Body-Mass-Index (BMI), where the body weight is related to the height of the person.

Consider a life insurance policy that is priced at two levels. The first level is an ordinary class rate where classification is made according to sharp criteria, and the second level is a proportional risk adjustment according to fuzzy variables. The sharp class rating is calculated in accordance with age and sex. To achieve further differentiation between applicants, insureds have to pay a risk loading λ on top of the class rate. Persons with a greater risk of having cardiovascular disease have to pay a positive risk loading ($\lambda > 0$), whereas persons with a lower risk pay a negative risk loading ($\lambda < 0$) - that is, healthy people get bonuses [5]. The risk loading is proportional to class rate and the gross premium is calculated as:

$$P(k/i,j) = C_{i,j} + \frac{\lambda_k}{100} C_{i,j} \ , \ i = age; j = sex \tag{7}$$

where $P(k/i,j)$ is the gross premium for person k with age i, sex j; $C_{i,j}$ is the class rate; and λ_k is the risk loading in a percentage. This pricing technique makes use of the relative strengths of both statistical inference and fuzzy inference. As statistics generates stable class rates with low variation for a limited number of risk factors, we classify on the first level according to the important characteristics age, sex. On the next level, we risk adjust the class rates according to fuzzy risk factors for propensity to cardiovascular disease. To evaluate the fuzzy risk factors and determine the risk loading λ, we employ ANFIS based system modeling. In this model, λ is an output of the system.

4 Numerical Application

In the proposed algorithm, before starting the process of solving, the class numbers of the independent variables are determined heuristically and are bound to numbers that are indicated at the beginning of the solution process. The fuzzy class numbers, which are determined intuitionally, are effective on a number of the models that are established in the solving process by being dependent to the variable numbers that occur in the model. The data set for independent variables and dependent variable is derived using simulation technique. In order to show the performance evaluation of the proposed method, the prediction value from this method and the errors related to these predictions are compared with other predictions and the errors, which are obtained from the Least Square Method (LSM).

The figures of errors obtained from two methods are given in Fig. 1. In Fig. 1, part [a] shows the errors related to the predictions that are obtained from the proposed algorithm and part [b] shows the errors related to the predictions that are obtained from The Least Squares Method.

For a better understanding of the main components of the proposed method, we consider a numerical example. The mortality risk due to a cardiovascular disease is predicted by the variables: $x_{1,k}$; $x_{2,k}$; $x_{3,k}$; $x_{4,k}$ where $x_{1,k}$ = systolic blood pressure, $x_{2,k}$ = cholesterol level, $x_{3,k}$ = body mass index, $x_{4,k}$ = the average cigarette consumption per day. Let $\hat{\lambda}_k$ denote the predicted risk loading in a percentage. Now, let us assume that prospective policyholder shows the following health

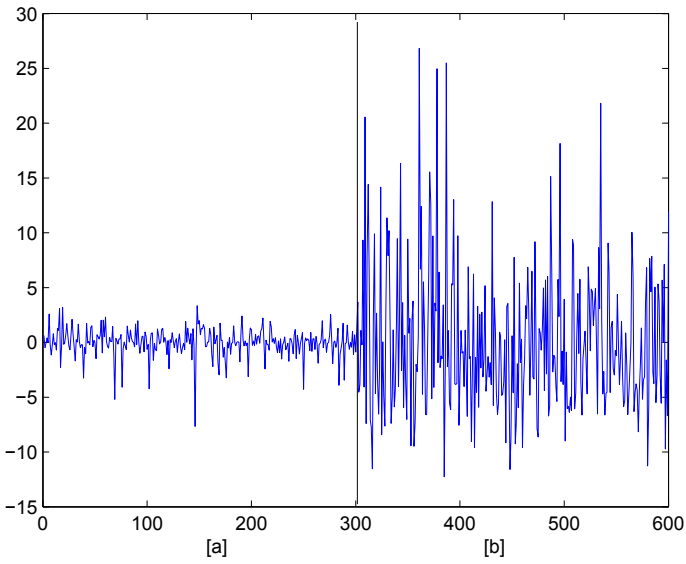


Fig. 1 Graphs of errors

profile:$x_1 = 143\ mm\ Hg$, $x_2 = 223\ mg\ /dl$, $x_3 = 26.1\ kg/m^2$, $x_4 = 4$. By examining the risk levels of these inputs, ANFIS infers the degree of risk loading an applicant must bear. The prediction value, related to the risk levels of these inputs are derived from the adaptive network as $\hat{\lambda}_k = 44.92$.

5 Conclusion

For risk classification in underwriting process, the uncertainty can be explained effectively with the fuzzy set concept which is accepted as a inference based on a specific logic and which helps in determining optimum decision at vagueness conditions. The process followed in the proposed method can be accepted as ascendant from other methods since it doesn't let the intuitional predictions and it brings us to the smallest error. At the same time; this method has the robust feature since it doesn't affected by the contradictory observations that can occur at independent variables. In the subsequent studies, this method can be compared with other robust methods.

References

1. C.B. Cheng and E.S. Lee. Applying Fuzzy Adaptive Network to Fuzzy Regression Analysis. *An International Journal Computers and Mathematics with Applications*, 38(2): 123–140, July 1999.
2. G.W. DeWit. Underwriting and Uncertainty. *Insurance: Mathematics and Economics*, 1(4): 277–285, October 1982.
3. T.D. Erbay and A. Apaydın. A Fuzzy Adaptive Network Approach to Parameter Estimation in Case where Independent Variables Come from Exponential Distribution. *Journal of Computational and Applied Mathematics*, 233(1): 36–45, November 2009.
4. T.A. Gaziano. Reducing the Growing Burden of Cardiovascular Disease in the Developing World. Health Affairs, 26(1): 13–24, January/February 2007.
5. P.J. Horgby. Risk Classification by Fuzzy Inference. *The Geneva Papers on Risk and Insurance*, 23(1): 63–82, June 1998.
6. H. Ishibuchi and H. Tanaka. Fuzzy Regression Analysis Using Neural Networks. *Fuzzy Sets and Systems*, 50(3): 257–265, September 1992.
7. J. Lemaire. Fuzzy Insurance. *Astin Bulletin*, 20(1): 33–56, April 1990.
8. J. Mackay and G. Mensah. *Atlas of Heart Disease and Stroke*. World Health Organization, Geneva 2004.
9. T. Takagi and M. Sugeno. Fuzzy Identification of Systems and its Applications to Modeling and Control. *IEEE Trans. on Systems, Man and Cybernetics*, 15(1): 116–132, January/Februray 1985.
10. V.R. Young. The Application of Fuzzy Sets to Group Health Underwriting. *Trans. Soc. Actuaries*, 45: 555–590, annually 1993.

Support Vector Regression for Time Series Analysis

Renato De Leone

Abstract In the recent years, Support Vector Machines (SVMs) have demonstrated their capability in solving classification and regression problems. SVMs are closely related to classical multilayer perceptron Neural Networks (NN). The main advantage of SVM is that their optimal weights can be obtained by solving a quadratic programming problem with linear constraints, and, therefore, standard, very efficient algorithms can be applied. In this paper we present a 0–1 mixed integer programming formulation for the financial index tracking problem. The model is based on the use of SVM for regression and feature selection, but the standard 2–norm of the vector w is replaced by the 1–norm and binary variables are introduced to impose that only a limited number of features are utilized. Computational results on standard benchmark instances of the index tracking problems demonstrate that good quality solution can be achieved in a limited amount of CPU time.

1 Support Vector Machine for Regression

Support Vector Machines (SVMs) are a novel and extremely effective tool for classification and regression [9, 5]. Let $\{(x^i, y_i)\}$, $i = 1, \ldots, l$ be the training data, where $x^i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, $i = 1, \ldots, l$; for the ε–Support Vector Regression problem [8], the aim is to determine a vector $w \in \mathbb{R}^n$ and a scalar b such that, for the function $f(x) = w^T x + b$, the deviation (with a tolerance of ε) of $f(x^i)$ from the corresponding target value y_i is as small as possible for each $i = 1, \ldots, l$. In addition, it is required the function $f(x)$ be as *flat* as possible, i.e., $\|w\|$ must be minimized. More specifically, the primal problem requires to solve the following quadratic optimization problem:

Renato De Leone
School of Science and Technology, Università di Camerino, via Madonna delle Carceri 9, Camerino (MC) ITALY, e-mail: renato.deleone@unicam.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_6,

$$\begin{array}{ll} \min\limits_{w,b,\xi^+,\xi^-} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{l}\left(\xi_i^+ + \xi_i^-\right) \\ \text{subject to} & w^T x^i + b - y_i \le \varepsilon + \xi_i^+, i = 1,\ldots,l \\ & w^T x^i + b - y_i \ge -\varepsilon - \xi_i^-, i = 1,\ldots,l \\ & \xi^+ \ge 0,\ \xi^- \ge 0 \end{array} \tag{1}$$

The dual of the above problem is

$$\begin{array}{ll} \min\limits_{\alpha^+,\alpha^-} & \frac{1}{2}\left(\alpha^+ - \alpha^-\right)^T Q\left(\alpha^+ - \alpha^-\right) + \varepsilon\sum\limits_{i=1}^{l}\left(\alpha_i^+ + \alpha_i^-\right) - \sum\limits_{i=1}^{l} y_i\left(\alpha_i^+ - \alpha_i^-\right) \\ \text{subject to} & \sum\limits_{i=1}^{l}\left(\alpha_i^+ - \alpha_i^-\right) = 0 \\ & \alpha_i^+ \in [0,C],\ \alpha_i^- \in [0,C], \quad i = 1,\ldots,l \end{array} \tag{2}$$

where $Q_{ij} := x^{iT} x^j$. From Karush–Kuhn–Tucker conditions [7], it follows that

$$w = \sum_{i=1}^{l}\left(\alpha_i^+ - \alpha_i^-\right) x^i$$

and, therefore, the function $f(x)$ assumes the form:

$$f(x) = \left(\sum_{h=1}^{l}\left(\alpha_h^+ - \alpha_h^-\right) x^h\right)^T x + b$$

Nonlinear classifiers can be obtained by applying the *kernel trick*. After defining a map $\psi : x \in \mathbb{R}^n \mapsto \psi(x) \in F$ where F ia an Hilbert space of finite or infinite dimension equipped with a scalar product $< .,. >$ in the new feature space, the training data are now $\left\{(\psi(x^i), y_i)\right\}, i = 1,\ldots,l$. For the dual problem (2) the only change needed is to replace $Q_{ij} := x^{iT} x^j$ with $\hat{Q}_{ij} :=< \psi(x^i), \psi(x^j) >$.

An important line of research has been devoted to the use of linear programming in solving the regression problem. The idea is to replace the 2–norm in (1) with the 1–norm. A first possibility is to use again the function $f(x) = w^T x + b$ and solve the following optimization problem

$$\begin{array}{ll} \min\limits_{w,b,\xi^+,\xi^-} & \sum\limits_{k=1}^{n}|w_k| + C\sum\limits_{i=1}^{l}\left(\xi_i^+ + \xi_i^-\right) \\ \text{subject to} & \sum\limits_{k=1}^{n} w_k x_k^i + b - y_i \le \varepsilon + \xi_i^+, i = 1,\ldots,l \\ & \sum\limits_{k=1}^{n} w_k x_k^i + b - y_i \ge -\varepsilon - \xi_i^-, i = 1,\ldots,l \\ & \xi^+ \ge 0, \xi^- \ge 0 \end{array} \tag{3}$$

and an equivalent linear programming problem can be obtained trough standard transformations.

A second possibility [10] is to use the Support Vector (SV) expansion of the vector w as a linear combination of the training patterns in the function $f(x)$:

$$f(x) = \left(\sum_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x^h \right)^T x + b$$

and the corresponding linear programming problem becomes:

$$\begin{array}{ll} \min\limits_{\alpha^+,\alpha^-,b,\xi^+,\xi^-} & \sum\limits_{i=1}^{l} \left(\alpha_i^+ + \alpha_i^- \right) + C \sum\limits_{i=1}^{l} \left(\xi_i^+ + \xi_i^- \right) \\ \text{subject to} & \left(\sum\limits_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x^h \right)^T x^i + b - y_i \leq \varepsilon + \xi_i^+, i = 1, \ldots, l \\ & \left(\sum\limits_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x^h \right)^T x^i + b - y_i \geq -\varepsilon - \xi_i^-, i = 1, \ldots, l \\ & \alpha^+ \geq 0,\ \alpha^- \geq 0,\ \xi^+ \geq 0,\ \xi^- \geq 0 \end{array} \tag{4}$$

Once again it is with notice that only scalar products ${x^h}^T x^i$ are involved and, therefore, it is possible to move to kernel functions. A similar approach in SVM classification has been proposed in [3]

2 Support Vector Regression for Feature Selection and Regression

In this section we consider a modification of the models proposed in the Section 1. The aim here is to determine, once again, a function $f(x) = w^T x + b$ but, in addition, we require that only a small percentage of the components w_k is different from 0. Also in this case, we can utilize the SV expansion of the vector w to obtain the following 0–1 mixed integer programming problem:

$$\begin{array}{ll} \min\limits_{\alpha^+,\alpha^-,b,\xi^+,\xi^-,\gamma} & \sum\limits_{i=1}^{l} \left(\alpha_i^+ + \alpha_i^- \right) + C \sum\limits_{i=1}^{l} \left(\xi_i^+ + \xi_i^- \right) \\ \text{subject to} & \left(\sum\limits_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x^h \right)^T x^i + b - y_i \leq \varepsilon + \xi_i^+, i = 1, \ldots, l \\ & \left(\sum\limits_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x^h \right)^T x^i + b - y_i \geq -\varepsilon - \xi_i^-, i = 1, \ldots, l \\ & \sum\limits_{h=1}^{l} \left(\alpha_h^+ - \alpha_h^- \right) x_k^h \leq M \gamma_k, \quad k = 1, \ldots, n \\ & \sum\limits_{k=1}^{n} \gamma_k \leq N \\ & \alpha^+ \geq 0,\ \alpha^- \geq 0,\ \xi^+ \geq 0,\ \xi^- \geq 0 \\ & \gamma_k \in \{0, 1\}, \quad k = 1, \ldots, n \end{array} \tag{5}$$

Note that, in this model, at most N components of the vector w will be different from 0. A possible alternative is to eliminate, in the model above, the constraint $\sum_{k=1}^{n} \gamma_k \leq N$ and add the term $\sum_{k=1}^{n} \gamma_k$ to the objective function with a suitable penalty coefficient.

3 Financial Index Tracking

Index tracking [4] is a form of passive fund management where the goal is to replicate (track) the performance of a market index using a limited number of the available stocks.

The values of n stocks, as well as the index to be traced are observed at time 0, 1, ..., T. At time T it must be decided the N ($N < n$) stocks to be acquired, as well as their quantities. Then, the predicted value is evaluated and compared to the observed value in the interval $T, \ldots, T'$, where $T' > T$. More specifically, let I_t be the value of the index to be traced and q_k^t the value of the stock k at time step t, where $t = 0, 1, \ldots, T$ and $k = 1, \ldots, n$. Moreover, a fixed amount of money B is available at time T to acquire the stocks. A in–depth literature review on index tracking can be found in [2, 4].

The model proposed is the following

$$
\begin{aligned}
\min_{\alpha^+,\alpha^-,b,\xi^+,\xi^-,\gamma} \quad & \sum_{t=0}^{T} \left(\alpha_t^+ + \alpha_t^-\right) + C \sum_{t=0}^{T} \left(\xi_t^+ + \xi_t^-\right) \\
\text{subject to} \quad & \sum_{k=1}^{n} \sum_{h=0}^{T} \left(\alpha_h^+ - \alpha_h^-\right) q_k^h q_k^t - \theta I_t \leq \varepsilon + \xi_t^+, \quad t = 0, \ldots, T \\
& \sum_{k=1}^{n} \sum_{h=0}^{T} \left(\alpha_h^+ - \alpha_h^-\right) q_k^h q_k^t - \theta I_t \geq -\varepsilon - \xi_t^-, \quad t = 0, \ldots, T \\
& 0 \leq \sum_{h=0}^{T} \left(\alpha_h^+ - \alpha_h^-\right) q_k^h \leq M\gamma_k, \quad k = 1, \ldots, n \qquad (*) \\
& \sum_{k=1}^{n} \gamma_k \leq N \qquad (*) \\
& \sum_{k=1}^{n} q_k^T \sum_{h=0}^{T} \left(\alpha_h^+ - \alpha_h^-\right) q_k^h = B \\
& \alpha^+ \geq 0,\ \alpha^- \geq 0,\ \xi^+ \geq 0,\ \xi^- \geq 0 \\
& \gamma_k \in \{0, 1\}, \qquad k = 1, \ldots, n
\end{aligned}
\tag{6}
$$

where the last constraint is the budget restriction constraint, the constraints (*) ensure that the chosen portfolio does not include more than N stocks and $\theta = B/I_T$.

4 Computational Results

The data set used in validating the model (6) proposed in the previous section are derived from the benchmark instances for index tracking in the OR–library maintained

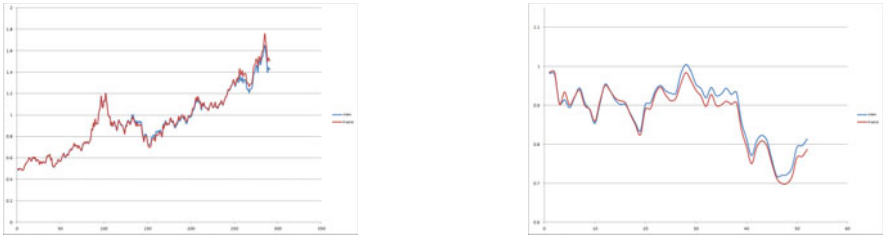
by Beasley [1]. An integer programming model for the index tracking problem was proposed also in [6].

The characteristics of the problems are the following

Table 1 Characteristics of the instances

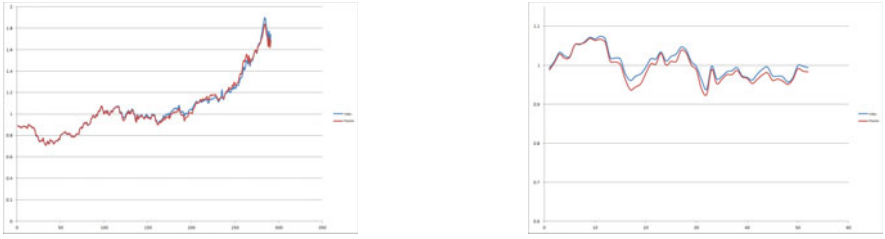
Instance	Market Index	n	T	N
indextracking1	Hang Seng	31	104	10
indextracking2	DAX 100	85	104	10
indextracking3	FTSE 100	89	104	10

We utilized MPL from Maximal Software to model our problem and CPLEX 12.1.0 and GUROBI 3.0 as solvers. All runs have been executed on a 64–bit Intel Core 2 Duo Mobile Processor T6600 (2.20GHz). The results obtained by both solvers are quite similar and, therefore, we report here the results obtained with GUROBI. For all problems we set the maximum CPU time to 1h.



(a) (b)

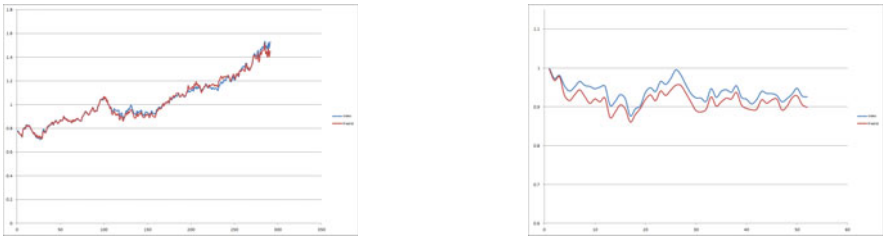
Fig. 1 Instance IndTrack1, Hang Seng



(a) (b)

Fig. 2 Instance IndTrack2, DAX 100

In all figures above, the index has been normalized to 1 at time T. The blue line is the value of the index while the red line is the value predicted by the model. The figure on the left (a) shows the behavior of the index to be traced and the calculated value over the entire period $[0, T' = T + 186]$. The figure (b) on the right,



(a) (b)

Fig. 3 Instance IndTrack3, FTSE 100

instead, reports the same values for the time period $[T, T+52]$. The results show a good predictive capability of the model, even far in the future. We refer to [6] for a comparison to the literature results for this class of problems.

These computational results show that Support Vector Machines can be successfully used in the financial index tracking problems. As future work, we plan to apply this technique to other time series prediction problems and to develop heuristic techniques to quickly obtain good quality solutions for the model (6).

References

1. J. E. Beasley. OR-library: Distributing test problems by electronic mail. *Journal of the Operational Research Society*, 41: 1069Ű-1072, 1990. http://dx.doi.org/10.1057/jors.1990.166.
2. J.E. Beasley, N. Meade, and T.-J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3): 621–643, August 2003.
3. K.P. Bennett. Combining support vector and mathematical programming methods for classification. In B. Schölkopf, C.J.C. Burges and A.J. Smola, editors, *Advances in kernel methods: support vector learning*, pages 307–326, MIT Press, Cambridge, MA, USA, 1999.
4. N.A. Canakgoz and J.E. Beasley. Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*, 196(1): 384–399, 2009.
5. N. Cristianini and J. Shave–Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
6. G. Guastaroba. *Portfolio Optimization: Scenario Generation, Models and Algorithms*. PhD thesis, Università degli Studi di Bergamo, 2010.
7. O.L. Mangasarian. *Nonlinear Programming*. McGraw–Hill, New York, 1969.
8. A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3): 199–222, 2004.
9. V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
10. J. Weston, A. Gammerman, M.O. Stitson, V.N. Vapnik, V. Vovk, and C. Watkins. Support vector density estimation. In B. Schölkopf, C.J.C. Burges and A.J. Smola, editors, *Advances in kernel methods: support vector learning*, pages 293–305. MIT Press, Cambridge, MA, USA, 1999.

I.2 Game Theory and Experimental Economics

Chair: Prof. Dr. Karl Morasch (Universität der Bundeswehr München)

Game Theory analyzes individual decision making in strategic situations. Unlike to a decision theoretic setting the outcome for an individual is then affected by actions of other players as well. Making sensible decisions therefore requires predicting the likely behavior of one's opponents. Assumptions about the rationality of the players and their knowledge of each other's preferences and available strategies are therefore central for the proper analysis of a game.

Most game theoretic results are quite sensitive with respect to small changes in the exact rules of the game (e.g. timing of moves, symmetry vs. asymmetry of information). Testing predictions derived from game theoretic models with field data is usually unfeasible as it is not possible to control for all these variations.

Experimental Economics can achieve the necessary level of control by testing theories in the lab where the rules of a game can be exactly replicated. Departures from the theoretically predicted results must then stem from inappropriate assumptions about the preferences or the rationality of the players.

We seek both papers that apply game theoretic modeling to any kind of management problem as well as experimental studies that test game theoretic results that are relevant for business, management and economics.

Intermediation by Heterogeneous Oligopolists

Karl Morasch

Abstract In his book on "Market Microstructure" Spulber presented some strange results with respect to the impact of the substitutability parameter in an intermediation model with differentiated products and inputs. Intuitively, effects in the product and the input market should be similar: if firms become more homogeneous, they loose market power, which should yield lower bid-ask-spreads and higher output. However, in Spulber's analysis parameter changes in the product market yield qualitatively different results for bid–ask spreads and output than equivalent changes in the input market. The present paper shows that this outcome stems from an inadequate normalization of demand in upstream and downstream markets, respectively. By appropriately controlling for market size effects, intuitive results are obtained. Beyond that, a setting with appropriate normalization also allows to address the impact of changes in the number of competitors on the market outcome.

1 Competition and Intermediation – The Intuition

Intermediaries establish and operate markets by buying from producing firms and selling to consumers. A monopolistic intermediary sets its bid and ask prices in a way to maximize profits. It behaves simultaneously as a monopolist in the output market and as a monopsonist in the input market. By equating marginal revenue to marginal expenditures the intermediary determines the optimal quantity q. Bid and ask prices, w and p, are then determined by the resulting prices on the supply and demand schedules, respectively. Figure 1 shows the market diagram with an monopolistic intermediary. Intermediation is a viable option if the intermediation rent as displayed in the figure is high enough to cover the cost of intermediation.

What happens if there are two competing intermediaries? Competition between intermediaries reduces the market power relative to the monopoly setting.

Karl Morasch
Universität der Bundeswehr München, e-mail: karl.morasch@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_7, © Springer-Verlag Berlin Heidelberg, 2011

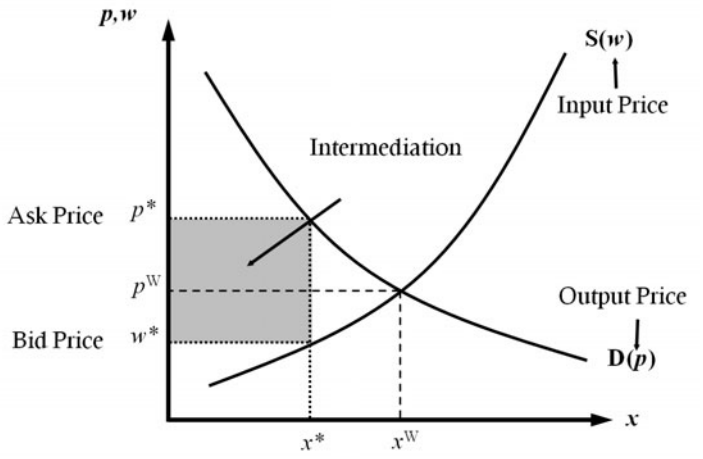


Fig. 1 Price setting by a monopolistic intermediary

Bid–ask spreads should be lower and the quantity of the two intermediaries together should be higher than the quantity of a single intermediary. With homogeneous goods and identical marginal costs of intermediation, both firms would choose a bid–ask spread that equals marginal costs of intermediation. A more realistic setting would introduce some kind of differentiation or switching costs between intermediaries. Considering the case with differentiation, the individual demand and supply schedules of the intermediaries should be flatter (more elastic demand and supply, respectively) than the market demand and supply curves. While customers of a monopolistic intermediary have only the option to refrain from buying or selling the good, respectively, they could now also switch to the competing intermediary. As shown in figure 2, bid–ask spreads are then lower and the total quantity is higher than in the monopoly case.

In his book on "Market Microstructure" [5] proposes a model with symmetric product differentiation where only two parameters – one for the demand side and the other for the supply side – determine the intensity of competition. Intuitively one would expect that bid–ask spreads would be reduced and output increased whenever intermediation services on either the demand or supply side become closer substitutes. However, in Spulber's analysis bid–ask spreads actually rise if the intermediaries become closer competitors on the demand side. Beyond that, both bid and ask prices rise with closer substitutes on the supply as well as on the demand side. How can these counterintuitive results be explained? In the following it is shown that this is due to not properly controlling for the market size effect of changes in the demand and supply parameters.

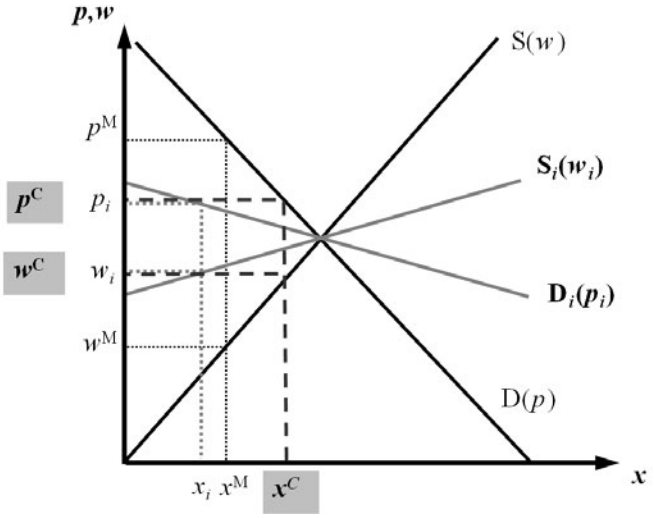


Fig. 2 Competing intermediaries with heterogeneous services

2 Spulber's Intermediation Duopoly

Spulber normalized marginal intermediation costs to zero and used the following seemingly sensible specification for demand and supply:

$$q_i = D_i(p_1, p_2) = 1 - p_i + t p_j, \tag{1}$$
$$x_i = S_i(w_1, w_2) = w_i - \tau w_j \tag{2}$$

with $i, j = 1, 2, i \neq j$, $0 < t < 1$, and $0 < \tau < 1$. If the parameters t and τ, respectively, take values close to zero, we approach independent intermediation services, if they take values close to one, the two services become almost perfect substitutes.

The two firms are assumed to choose output and input prices simultaneously. In an equilibrium $(w_1^*, p_1^*, w_2^*, p_2^*)$ firm 1 chooses prices (p_1^*, w_1^*) that maximize profits

$$\pi_1(w_1, p_1, w_2^*, p_2^*) = p_1 D_1(p_1, p_2^*) - w_1 S_1(w_1, w_2^*) \tag{3}$$
$$\text{such that} \quad S_1(w_1, w_2^*) \geq D_1(p_1, p_2^*) \tag{4}$$

where (p_2^*, w_2^*) indicates the equilibrium prices of firm 2. Due to symmetry firm two's problem is similar.

In equilibrium the stock constraint is binding and we obtain the following prices and the resulting outputs (the index is skipped as prices will be identical in a symmetric equilibrium):

$$p^* = \frac{3-2\tau}{4-3(t+\tau)+2t\tau} \tag{5}$$

$$w^* = \frac{1}{4-3(t+\tau)+2t\tau} \tag{6}$$

$$q^* = \frac{1-\tau}{4-3(t+\tau)+2t\tau} \tag{7}$$

Looking at the equation for q^* one immediately notices that changes in substitutability in the input market have different impact than similar changes in the output market. Taking a closer look, one observes the peculiar result that all prices rise if intermediation services become closer substitutes in either output or input markets. However, while bid–ask spreads increase with rising t, they decrease when τ becomes larger.

What is the reason for these counterintuitive results? Changes in t and τ affect the degree of substitutability (the point we are interested in), but they also affect market size: For a given value of p_j the demand curve shifts outward by tp_j relative to the monopoly demand curve for $t = 0$ (market size increases). In the same manner the supply curve is shifted inward by a rise in τ (decreasing market size).

3 Proper Normalization with a Love of Variety Approach

As an alternative to Spulber's specification, we propose to start from a system of inverse demand and supply that is properly rooted in a utility maximization and cost minimization problem. To save space we will only explicitly consider the demand side that is based on the love of variety approach of product differentiation pioneered by [4] and [1]. Here the consumption side is given by a representative consumer with linear-quadratic utility

$$U(q_1, q_2; q_0) = \alpha(q_1 + q_2) - \frac{1}{2}(q_1^2 + q_2^2 + 2bq_1q_2) + q_0 \tag{8}$$

with q_1 and q_2 indicating the specific types of the differentiated good produced by firm 1 or 2, respectively, and q_0 a numeraire good which is assumed to be produced in another sector of the economy and has been added linearly to ensure that the marginal utility of income is equal to one. The parameter α is a measure of market size while b describes the degree of substitutability between the products of the two firms: If the products are perfect substitutes $b = 1$, if they are independent $b = 0$. For the ease of computation and to show the similarity to Spulber's setting, the market size parameter is normalized to $\alpha = 1$.

Given the utility function for $\alpha = 1$, the consumer maximization problem leads to linear inverse demand functions

$$p_i = 1 - q_i - bq_j \quad \text{with} \quad j \neq i. \tag{9}$$

Demand functions expressing quantity demanded as a function of the two prices are necessary to discuss intermediation. Based on the two inverse demand functions straightforward calculations yield

$$D_i(p_1, p_2) = \frac{1}{1-b^2}[(1-b) - p_i + bp_j]. \tag{10}$$

Note that this demand function only differs from the one used by Spulber with respect to the multiplicative term $1/(1-b)$ and the normalized intercept $(1-b)$. Applying similar reasoning on the supply side, we obtain supply functions

$$S_i(w_1, w_2) = \frac{1}{1-\beta^2}[w_i - \beta w_j]. \tag{11}$$

The resulting prices and quantities in equilibrium are then given by the following expressions:

$$p^* = \frac{3-b^2+\beta(1-\beta)}{4+b(1-b)+\beta(1-\beta)} \tag{12}$$

$$w^* = \frac{1+\tau}{4+b(1-b)+\beta(1-\beta)} \tag{13}$$

$$q^* = \frac{1}{4+b(1-b)+\beta(1-\beta)} \tag{14}$$

While the numerators for p^* and w^* still differ in a way that it is not immediately obvious that impacts of changes in supply and demand are similar, the formula for the equilibrium quantity q^* is perfectly symmetric. This implies that the same must be true for the bid–ask spread and by subtracting w^* from p^* and simplifying appropriately we actually obtain

$$p^* - w^* = \frac{2-b^2-\beta^2}{4+b(1-b)+\beta(1-\beta)}. \tag{15}$$

Closer inspection shows that the bid–ask spread is indeed reduced whenever intermediary services become closer substitutes on the demand or the supply side.

A sensibly specified model should also behave reasonably at the limits of the parameter space. The Cournot oligopoly for example approaches the solution under perfect competition if the number of firms gets very large. For this reason we check what happens if b and β both simultaneously approach the limiting values of zero and one, respectively.

- For $b, \beta \to 0$ the services become independent and, as should be the case, we obtain the monopolistic intermediary solution for linear demand $D_i(p_i) = 1 - q_i$, namely $p_i^* = 3/4$, $w_i^* = 1{,}4$ and $q_i^* = 1/4$.
- In a similar manner $b, \beta \to 1$ yields the Walras equilibrium without any intermediary rents, i. e. $p_i^* = w_i^* = 1/2$ and $q_i^* = 1/2$.

4 Conclusion and Extensions

While our approach yields reasonable and "well behaved" results, it should be noted, that market size is also changing here (but in a more sensible manner than in Spulber's analysis). Due to the assumption of love of variety the market with independent services is twice as large (two monopoly markets) than the market with homogeneous services (just one market with two firms). Beyond that a larger number of firms than two would also increase the market size in this setting.

There exists another differentiated products model due to [3] (see also [2] for applications in oligopoly theory and competition policy). This model is more complicated but still analytically tractable and has the nice property that aggregate demand does neither depend on the degree of substitution among products or services nor on the number of firms. As expected, using this kind of model yields qualitatively similar results with respect to bid–ask spread and quantity in the duopoly setting. Because it is not very reasonable to assume that demand and supply rises with the number of intermediaries, the approach by Shubik and Levitan seems to be preferable for extending the analysis to the oligopoly setting.

References

1. A. Dixit and J. Stiglitz. Monopolistic Competition and Optimum Product Diversity. *American Economic Review*, 67: 287–308, , 1977.
2. M. Motta. *Competition Policy. Theory and Practice*. Cambridge University Press, Cambridge (UK), 2004.
3. M. Shubik and R. Levitan. *Market Structure and Behavior*. Harvard University Press, Cambridge Cambridge (MA), 1980.
4. M. Spence. Product Selection, Fixed Costs and Monopolistic Competition. *Review of Economic Studies*, 43: 217–235, 1976.
5. D. F. Spulber. *Market Microstructure. Intermediaries and the Theory of the Firm*. Cambridge University Press, Cambridge (UK), 1999.

Algorithmic Aspects of Equilibria of Stable Marriage Model with Complete Preference Lists

Tomomi Matsui

Abstract Given two sets of agents, men and women, Gale and Shapley discussed a model, called the *stable marriage model*, in which each agent has a preference over agents of the opposite sex. Gale and Shapley showed that every set of preference lists admits at least one stable marriage by describing an algorithm, called the Gale-Shapley algorithm, which *always* finds a stable marriage.

Given (true) preference lists of men over women and (true) preference lists of women over men, we introduce a game among women. In a play of the game, each woman chooses a strategy which corresponds to a complete preference list over men. The resulting payoff of a woman is her mate determined by men-proposing Gale-Shapley algorithm executed on men's (true) preference lists and women's joint strategy. We propose a polynomial time algorithm for checking whether a given marriage is an equilibrium outcome or not.

1 Introduction

Given two sets of agents, men and women, Gale and Shapley [2] discussed a model, called the *stable marriage model*, in which each agent has a preference over agents of the opposite sex. A stable marriage is a one-to-one mapping between sets of men and women, such that there is no man-woman pair who would agree to leave their assigned mates in order to marry each other. Gale and Shapley showed that every set of preference lists admits at least one stable marriage by describing an algorithm, called the deferred acceptance algorithm (the Gale-Shapley algorithm), which *always* finds a stable marriage.

This paper deals with a strategic issue in the stable marriage model with complete preference lists. Given (true) preference lists of men over women and (true) prefer-

Tomomi Matsui
Department of Information and System Engineering, Faculty of Science and Engineering, Chuo University, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_8,

ence lists of women over men, we introduce a game among women. In a play of the game, each woman chooses a strategy which is a complete preference list over men. The resulting payoff of a woman is her mate determined by men-proposing Gale-Shapley algorithm executed on men's (true) preference lists and women's joint strategy. We propose a polynomial time algorithm for checking whether a given marriage is an equilibrium outcome or not.

The issues of strategic manipulation in the stable marriage context are discussed in some papers (see [5, 10] for example). Roth [8] showed that when the men-proposing algorithm is used, none of the men benefits by submitting a false preference list. Dubins and Freedman [1] proved that no coalition of men could collectively manipulate their preferences in such a way as to strictly improve all of their mates. In settings that allow incomplete preference lists, women on the other hand have incentives to cheat in the men-proposing algorithm. Gale and Sotomayor [3] showed that a woman has an incentive to falsify her preference as long as she has at least two distinct stable mates.

In 1991, Zhou [12] gave a characterization of equilibrium outcomes (marriages). His result does not give an efficient algorithm for checking whether a given marriage is an equilibrium outcome or not. Teo, Sethuraman, and Tan [11] deal with the situation where there exists a specified woman w who is the only deceitful agent. They proposed a polynomial time algorithm for constructing woman w's optimal cheating strategy. In recent papers [6, 7], author discussed problems for constructing a set of women's complete preference lists resulting a given (partial) marriage. Main results obtained in this paper employ some algorithms proposed in these papers.

2 Notations and Definitions

We denote two disjoint sets of agents by M and W, called men and women. A *marriage* is a mapping $\mu : (M \cup W) \to (M \cup W)$ satisfying

1. $\forall m \in M,\ \mu(m) \in W \cup \{m\}$,
2. $\forall w \in W,\ \mu(w) \in M \cup \{w\}$,
3. $i = \mu(j)$ if and only if $j = \mu(i)$ and
4. $|\{i \in M \cup W \mid \mu(i) \neq i\}| = 2\min\{|M|, |W|\}$.

If an agent $i \in M \cup W$ satisfies $\mu(i) \neq i$, we say that i is *matched* and $\mu(i)$ is the *mate* of i in marriage μ. We say that a pair of matched agents $\{m, \mu(m)\}$ (satisfying $m \neq \mu(m)$) is a *pair of mates*. Every agent with $\mu(i) = i$ is called *unmatched*. Obviously, the number of unmatched agents is equal to $||M| - |W||$.

Each agent in $M \cup W$ has a *preference list* which is a totally ordered list of all the members of the opposite sex. Agent i prefers q to r if and only if (1) q precedes r on i's preference list, or (2) agent q is identical with agent r. When i prefers q to r and q differs form r, we say that i strictly prefers q to r. We denote a set of true preference lists of M over W by $L^{(M)}$. Similarly, $L^{(W)}$ denotes a set of true preference lists of W

over M. A pair $(m,w) \in M \times W$ is called a *blocking pair* of a marriage μ, if at least one of the following conditions are hold;

1. both m and w are matched in μ, m strictly prefers w to $\mu(m)$ and w strictly prefers m to $\mu(w)$,
2. m is unmatched, w is matched, and w strictly prefers m to $\mu(w)$,
3. m is matched, w is unmatched, and m strictly prefers w to $\mu(m)$.

A marriage with no blocking pair is called a *stable marriage*. Gale and Shapley [2] showed that a stable marriage always exists, and a simple algorithm called the *deferred acceptance algorithm* can find a stable marriage. In the rest of this paper, the men-proposing deferred acceptance algorithm is called the *GS-algorithm*. We denote a marriage obtained by the GS-algorithm applied to a pair of lists $(L^{(M)}, L^{(W)})$ by $\mathrm{GS}(L^{(M)}, L^{(W)})$. It is well-known that the GS-algorithm produces the men-optimal marriage [5].

3 The Model and Main Result

We introduce a game defined by the true preference lists $(L^{(M)}, L^{(W)})$. We put the set of players to W. Every player (woman) has a common strategy set $\mathcal{P}$ which is a set of all the totally ordered lists of men in M. For simplicity, we denote the set of joint strategies $\times_{w \in W} \mathcal{P}$ by $\mathcal{P}^W$. A joint strategy in $\mathcal{P}^W$ corresponds to a set of preference lists of W over M. Given a joint strategy $K \in \mathcal{P}^W$, the payoff of a player $w \in W$ is her mate $\mu(w)$ where $\mu = \mathrm{GS}(L^{(M)}, K)$. For any pair of marriage (μ, μ') and a player (woman) $w \in W$, we say that w prefers μ to μ', denoted by $\mu \succeq_w \mu'$ if and only if

1. w is unmatched in μ', or
2. w is matched in both μ and μ', and w prefers $\mu(w)$ to $\mu'(w)$ with respect to her true preference list in $L^{(W)}$.

Throughout this paper, the relation $\succeq_w$ is defined with respect to the 'true' preference lists $L^{(W)}$. When both $\mu \succeq_w \mu'$ and $\mu(w) \neq \mu'(w)$ holds, we denote $\mu \succ_w \mu'$.

Given a joint strategy $K \in \mathcal{P}^W$, a (woman) player $w \in W$, and a totally ordered list $\pi \in \mathcal{P}$, a doublet (K_{-w}, π) denotes a joint strategy obtained from K by substituting π for w's preference list. A joint strategy $K \in \mathcal{P}^W$ is called an equilibrium point, when K satisfies that

$$\forall w \in W,\ \forall \pi \in \mathcal{P},\ \mu(w) \succeq_w \mu'(m),$$
$$\text{where } \mu = \mathrm{GS}(L^{(M)}, K) \text{ and } \mu' = \mathrm{GS}(L^{(M)}, (K_{-w}, \pi)).$$

When $|M| = |W|$, Zhou [12] showed that a marriage μ is an outcome of an equilibrium point if and only if μ is stable and attainable. It is easy to extend his result and obtain the following theorem.

Theorem 1 *For any marriage μ, there exists an equilibrium point $K \in \mathcal{P}^W$ satisfying $\mu = \mathrm{GS}(L^{(M)}, K)$ if and only if μ satisfies the conditions that*
(C1) *μ is stable with respect to the true preference lists $(L^{(M)}, L^{(W)})$, and*
(C2) *there exist a set of preference lists $\widetilde{K} \in \mathcal{P}^W$ with $\mu = \mathrm{GS}(L^{(M)}, \widetilde{K})$.*
(It is possible that $\widetilde{K}$ in (C2) differs from K.)

When $|M| > |W|$, we have the following,

Corollary 1 *If $|M| > |W|$, a marriage μ has an equilibrium point $K \in \mathcal{P}^W$ satisfying $\mu = \mathrm{GS}(L^{(M)}, K)$ if and only if μ is stable with respect to $(L^{(M)}, L^{(W)})$.*

Proof. We only need to show that μ satisfies Condition (C2). We construct $\widetilde{K} \in \mathcal{P}^W$ satisfying (1) the first choice of each woman is her mate in μ and (2) there exists an unmatched man $m \in M$ with respect to μ who is the second choice of every woman. It is easy to show that $\mu = \mathrm{GS}(L^{(M)}, \widetilde{K})$. □

The purpose of this paper is to prove the following theorem.

Theorem 2 *Let μ be a given marriage. There exists an $\mathrm{O}(|M||W|)$ time algorithm for checking the existence of an equilibrium point $K \in \mathcal{P}^W$ satisfying $\mu = \mathrm{GS}(L^{(M)}, K)$.*

4 Decision Problem of Equilibrium Outcomes

In this section, we give a proof of Theorem 2.
Proof of Theorem 2. Theorem 1 implies that we only need to check Conditions (C1) and (C2). We can check Condition (C1) (the stability of a given marriage μ with respect to the true preference lists $(L^{(M)}, L^{(W)})$) in $\mathrm{O}(|M||W|)$ time. In the following, we assume that μ is stable with respect to $(L^{(M)}, L^{(W)})$.

If $|M| > |W|$, the proof of Corollary 1 implies that we can construct a joint strategy $\widetilde{K}$ satisfying $\mu = \mathrm{GS}(L^{(M)}, \widetilde{K})$ in $\mathrm{O}(|M||W|)$ time.

Lastly, we consider the case that $|M| \leq |W|$. Let W^* be a set of matched women in μ. For each man $m \in M$, we construct a preference list from his preference list in $L^{(M)}$ by removing all women not in W^*. Denote the obtained set of lists of M over W^* by $L_*^{(M)}$. We show that we only need to check the existence of a set of preference lists K^* of women W^* over men M satisfying that the set of pairs of mates in $\mathrm{GS}(L_*^{(M)}, K^*)$ is equivalent to that in μ.

Assume that there exists a joint strategy $\widetilde{K} \in \mathcal{P}^W$ satisfying $\mu = \mathrm{GS}(L^{(M)}, \widetilde{K})$. Because μ is stable with respect of $(L^{(M)}, L^{(W)})$, every unmatched woman w in μ satisfies that each man m prefers $\mu(m)$ to w with respect to $L^{(M)}$ and thus no man proposed to w when we applied the GS-algorithm to $(L^{(M)}, \widetilde{K})$ (not to $(L^{(M)}, L^{(W)})$). Let $K^* \in \mathcal{P}^{W^*}$ be a subset of $\widetilde{K}$ consisting of preference lists of women in W^*. From the definition of the GS-algorithm, it is easy to see that the set of pairs of mates in $\mathrm{GS}(L_*^{(M)}, K^*)$ is equivalent to that in μ.

Conversely, assume that there exists a set of preference lists $K^* \in \mathcal{P}^{W^*}$ satisfying that the set of pairs of mates in $\mathrm{GS}(L_*^{(M)}, K^*)$ is equivalent to that in μ. Let $\widetilde{K} \in \mathcal{P}^W$ be a joint strategy obtained from K^* by adding an arbitrary preference list in $\mathcal{P}$ for each woman not in W^*. When we apply the GS-algorithm to $(L^{(M)}, \widetilde{K})$, the stability of μ with respect to $(L^{(M)}, L^{(W)})$ implies that each woman $w \notin W^*$ is proposed by no man. Thus, the resulting marriage $\mathrm{GS}(L^{(M)}, \widetilde{K})$ becomes μ.

From the above discussion, there exists a joint strategy $\widetilde{K} \in \mathcal{P}^W$ satisfying $\mu = \mathrm{GS}(L^{(M)}, \widetilde{K})$ if and only if there exists a set of preference lists $K^* \in \mathcal{P}^{W^*}$ satisfying that the set of pairs of mates in $\mathrm{GS}(L_*^{(M)}, K^*)$ is equivalent to that in μ. Thus, we only need to solve the following problem.

Problem $\mathrm{Q1}(L_*^{(M)}, \rho)$:

Input: A pair of sets (M, W^*) satisfying $|M| = |W^*|$, set of preference lists $L_*^{(M)}$ of men M over women W^* and a (perfect) marriage ρ defined on $M \cup W^*$.

Question: If there exists a set of preference lists K^* of women W^* over men M such that $\mathrm{GS}(L_*^{(M)}, K^*) = \rho$, then output K^*. If not, say 'none exists.'

In papers [6, 7], author proposed an $\mathrm{O}(|M||W^*|)$ time algorithm for Problem $\mathrm{Q1}(L_*^{(M)}, \rho)$.

Summarizing the above, we have the following algorithm. First, we construct the set W^* of matched women in μ, the set of preference lists $L_*^{(M)}$ obtained from $L^{(M)}$ by restricting to W^*, and a marriage ρ on $M \cup W^*$ consisting of pairs of mates in μ, which requires $\mathrm{O}(|M||W|)$ time. Next, we solve Problem $\mathrm{Q1}(L_*^{(M)}, \rho)$. If the solution of $\mathrm{Q1}(L_*^{(M)}, \rho)$ is 'none exists', there does not exists a joint strategy $\widetilde{K}$ satisfying Condition (C2). When our algorithm for Problem $\mathrm{Q1}(L_*^{(M)}, \rho)$ outputs a preference list $K^* \in \mathcal{P}^{W^*}$, we construct a joint strategy $\widetilde{K} \in \mathcal{P}^W$ from K^* by adding an arbitrary preference list in $\mathcal{P}$ for each woman not in W^*. Then $\widetilde{K}$ satisfies Condition (C2). The total computational time of this algorithm is bounded by $\mathrm{O}(|M||W|)$. □

We briefly describe our result on Problem $\mathrm{Q1}(L_*^{(M)}, \rho)$ appearing in [7]. We say that man m is a *suitor* of $w \in W^*$ if and only if (1) $w = \rho(m)$ or (2) m prefers w to his mate $\rho(m)$. A *rooted suitor graph*, denoted by $\overline{G}(L_*^{(M)}, \rho)$, is a directed bipartite graph with a vertex set $M \cup W^* \cup \{r\}$, where r is an artificial vertex called the *root*, and a set of directed edges A defined by

$$\begin{aligned} A = &\{(w, \rho(w)) \in W^* \times M \mid w \in W^*\} \\ &\cup \{(m, w) \in M \times W^* \mid m \text{ strictly prefers } w \text{ to } \rho(m)\} \\ &\cup \{(r, w) \mid w \text{ has a unique suitor } \rho(w)\}. \end{aligned}$$

Theorem 3 [7] *Problem* $\mathrm{Q1}(L_*^{(M)}, \rho)$ *has a solution (joint strategy)* $K^* \in \mathcal{P}^{W^*}$ *satisfying* $\mathrm{GS}(L_*^{(M)}, K^*) = \rho$ *if and only if the rooted suitor graph* $\overline{G}(L_*^{(M)}, \rho)$ *has an outgoing spanning tree with the root* r.

From the above theorem, we only need to search the rooted suitor graph $\overline{G}(L_*^{(M)}, \rho)$ for solving Problem Q1$(L_*^{(M)}, \rho)$.

5 Discussion

In this paper, we gave a algorithmic characterization of an equilibrium outcome. A remained important issue is the problem for checking whether a given joint strategy is an equilibrium point or not. For this problem, results in [7] imply the following.

Theorem 4 *There exists an* $O(|M|^2|W|^2)$ *time algorithm for checking whether a given joint strategy* $K \in \mathcal{P}^W$ *is an equilibrium point or not.*

References

1. L. E. Dubins and D. A. Freedman, "Machiavelli and the Gale-Shapley Algorithm," *American Mathematical Monthly*, vol. 88, pp. 485–494, 1981.
2. D. Gale and L. S. Shapley, "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, vol. 69, pp. 9–15, 1962.
3. D. Gale and M. Sotomayor, "Some Remarks on the Stable Matching Problem," *Discrete Applied Mathematics*, vol. 11, pp. 223–232, 1985.
4. D. Gale and M. Sotomayor, "Ms Machiavelli and the Stable Matching Problem," *American Mathematical Monthly*, vol. 92, pp. 261–268, 1985.
5. D. Gusfield and R. W. Irving, *The Stable Marriage Problem: Structure and Algorithms*, MIT Press, Cambridge, MA, 1989.
6. H. Kobayashi and T. Matsui, "Successful Manipulation in Stable Marriage Model with Complete Preference Lists," IEICE Trans. Inf. Syst., vol. E92-D, pp. 116–119, 2009.
7. H. Kobayashi and T. Matsui, "Cheating Strategies for the Gale–Shapley Algorithm," Algorithmica, vol. 58, pp. 151–169, 2010.
8. A. E. Roth, "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, vol. 7, pp. 617–628, 1982.
9. A. E. Roth, "Misrepresentation and Stability in the Stable Marriage Problem," *Journal of Economic Theory*, vol. 34, pp. 383–387, 1984.
10. A. E. Roth and M. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Cambridge, Cambridge University Press, 1990.
11. C.-P. Teo, J. Sethuraman, and W.-P. Tan, "Gale-Shapley Stable Marriage Problem Revisited: Strategic Issues and Applications," *Management Science*, vol. 47, pp. 1252–1267, 2001.
12. L. Zhou, "Stable Matchings and Equilibrium Outcomes of the Gale-Shapley's Algorithm for the Marriage Problem," *Economics Letters*, vol. 36, pp. 25–29.

Centralized Super-Efficiency and Yardstick Competition – Incentives in Decentralized Organizations

Armin Varmaz, Andreas Varwig and Thorsten Poddig

Abstract Due to deficient instruments of performance management, decentralized organizations often produce inefficiently. Promising approaches to dynamic incentives and performance management have recently been developed based on Data Envelopment Analysis (DEA). These, however, are not yet able to account for the specific needs of central performance management. We develop two new intra-organizational performance measures for defining dynamic incentive schemes and increasing overall performance. For suggestive evidence we evaluate the performances of 11 bank branches.

1 Introduction

Among other reasons, due to asymmetric information and individual ambitions of semi-autonomous decision makers, the production of goods and/or services in complexly structured organizations often is inefficient. This problem is particularly evident in decentralized organizations such as banks. Putting own interests first, single branches and branch managers often fight for larger fractions of shared resources, thereby impeding the efforts made by a central management to enhance the organization's overall performance. Such intra-organizational *market failure* can lead to a suboptimal allocation of resources and cause avoidable inefficiencies.
[3, 4] has developed promising approaches to dynamic incentive schemes and performance management. Introducing DEA based Yardstick Competition (DBYC), he demonstrates how relative performance estimation can be used to induce and con-

Armin Varmaz
University of Freiburg, Department of Finance

Andreas Varwig
University of Bremen, Department of Finance, e-mail: varwig@uni-bremen.de

Thorsten Poddig
University of Bremen, Department of Finance

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_9,

trol competition on non-competitive markets. However, the incentive mechanisms developed so far aim only on individual performance maximization. The consequences of individual behavior modifications for other branches of the same system are thereby mostly neglected. Also the possible benefits from an internal reallocation of resources are not considered. For managing performances within decentralized organizations we hence recommend to change the perspective of performance evaluation. Based on a DEA model for centralized resource planning by [6] we develop two new intra-organizational performance measures that can be used to define dynamic incentive schemes for branches and branch managers.

2 DBYC and Centralized Super-Efficiencies

The work of [3] on DEA based incentives originates from the Yardstick Competition (YC) model of [7], which is a game theoretical approach, enabling a regulator to control the behavior of non-competing firms. This idea can easily be transfered to intra-organizational performance management. By comparing similar organizations, efficient cost levels, which can be used to set up performance-related payment schemes, are identified. To central managers these levels serve as effective instruments to induce and control competition within the organization and achieve cost reductions. Despites possibly asymmetric information between the central manager (principal) and the branches (agents) the problems of *moral hazard* and *adverse selection* can be avoided and organizational performance be enhanced. Depending on a branch's individual performance, its production costs are proportionately refunded by the central management. While poor performers are not fully reimbursed, the costs of good performers are overcompensated. As a result, branches have an incentive for permanent process enhancements and cost reductions. The resulting optimal payment scheme is shown in (1).

$$b_i^* = X_i + \rho_i[\theta_i - 1]X_i, \qquad \forall i. \tag{1}$$

Transfers (b_i^*) to be paid to a branch (i) (managers and/or employees) depend on the branch's individual costs (X_i), its performance (θ_i) and a parameter ρ_i. The latter has to be negotiated between central management and the branches and defines a fraction of costs to be relevant for the calculation of rewards and penalties, respectively. Poor performers are thereby guaranteed a basic income that prevents from immediate failure and provides time for adjustments.[1] While [7] computes the individual θ_i as ratio of the optimal cost-level c_i^{**}, calculated as the average costs of all other branches, and the costs that have been disclosed by branch i, [3] calculates c_i^{**} based

[1] Note that (1) is appropriate only to a limited extent as a permanent incentive. In the long run this payment scheme could lead to a concentration of resources on few or even a single, well performing branch. Consequently, many underperforming branches would have to be closed. This is not always intended by the central management. In such situations a payment scheme based on performance rankings, perhaps particularly dealing with top performers, seems reasonable.

on DEA. DEA enables to determine single, unambiguous measures of efficiency and allows to evaluate large universes of organizations. By means of linear programming, best practices and adequate benchmarks for the inefficient branches are identified. However, as can be seen from (1), the performance estimator θ_i has to be able to take values above 1. If this requirement is not met, agents would only receive negative incentives, i.e. punishments for performing worse than best practice. Consequently, they would only try to perform as good as best practice, but would have no incentives for further improvements. Hence, the use of ordinary DEA-models, where θ_i is bounded between 0 and 1, would be inappropriate. One solution to this problem are *super-efficiencies*. Measures of super-efficiency can take values above 1 and calculate the degree to which efficient branches outperform the inefficient.[2] Based on the super-efficiency-model of [1], [3] develops a modified approach which allows external information to be taken into account.

Since the introduction of DEA-based regulatory mechanisms, their applicability and utility have been analyzed in many studies. Applications, however, are mostly based on classic DEA models and focus on inter-organizational performance comparisons. Although enabling to derive incentives that help to achieve individually minimal costs, ordinary DEA-models cannot account for shared resources and other interdependencies between the branches. Furthermore, they aim to optimize only individual production plans. Possible improvements of an organization's overall performance by reallocating resources among the branches hence remain unconsidered. Another drawback is the possible occurrence of hyper-efficiencies. Under certain conditions the linear programs for super-efficiencies become infeasible. This renders a complete evaluation of performances impossible and complicates the definition of incentive schemes.

Until today the issue of overall performance management has been addressed by various modifications and extensions of the basic DEA-models. A broad overview can be found at [5]. One approach that explicitly supports centralized resource planning and takes an internal perspective on performance evaluation has been proposed by [6], CRA-DEA. In contrast to inter-organizational DEA models, CRA-DEA simultaneously optimizes the performances of all organizations under evaluation. Thus, to analyze a network of n branches, instead of solving n optimization problems, only one linear program has to be solved. Considering possible improvements by the reallocation resources among all branches, an optimal production structure of the entire organization is computed. As mentioned earlier, CRA-DEA only computes one single measure of efficiency for an entire organization, $\tilde{\theta}$.[3] To derive a CRA-DEA based incentive scheme, however, individual performance measures are needed. For this task we subsequently develop two approaches.

Centralized super-efficiencies can be calculated in a two-step procedure. Therefore, the model of [6] has to be solved $n+1$ times, where n denotes the number of branches to be evaluated. Firstly $\tilde{\theta}$ is computed. The subsequent n optimizations

[2] An introduction to the functionality and applicability of super-efficiencies can be found at [8].

[3] Note that $\tilde{\theta}$ denotes the system's overall efficiency, assuming that all inputs are reduced proportionally. In case a non-radial CRA-DEA approach is used, efficiency is calculated for each input factor.

are carried out with a smaller dataset. In each optimization the dataset is reduced by the information on one DMU at a time. Thus, n datasets are formed, each of which contains data on the k inputs (X) and the r outputs (Y) of $n-1$ DMUs. The relating linear program is shown in (3), whereas the # denotes the reduced dataset.

$$\begin{aligned} \min \quad & \tilde{\theta}^{\#} \\ s.t. \quad & \sum_{j=1}^{n-1}\sum_{i=1}^{n-1} \tilde{\lambda}_{ji} x_{ki}^{\#} \leq \tilde{\theta}^{\#} \sum_{i=1}^{n-1} x_{ki}^{\#}, \quad \forall k \\ & \sum_{j=1}^{n-1}\sum_{i=1}^{n-1} \tilde{\lambda}_{ji} y_{ri}^{\#} \geq \sum_{i=1}^{n-1} y_{ri}^{\#}, \quad \forall r \\ & \sum_{i=1}^{n-1} \tilde{\lambda}_{ij} = 1, \quad \forall j \\ & \tilde{\lambda}_{ij} \geq 0; \quad \tilde{\theta}^{\#} \; free \end{aligned} \tag{2}$$

Let $\hat{\tilde{\theta}}_i^{\#}$ denote the n solutions to the program. These efficiency parameters represent the organization's overall performance, assuming that one branch i does not exist. The contribution of this branch to overall performance can now easily be calculated by the ratio of overall and partial performance, see equation (3).

$$\tilde{\theta}_i^a = \frac{\tilde{\theta}}{\hat{\tilde{\theta}}_i^{\#}} \qquad \text{and} \qquad \tilde{\theta}_i^b = \frac{(1-\hat{\tilde{\theta}}_i^{\#})}{(1-\tilde{\theta})} \tag{3}$$

The parameter $\tilde{\theta}_i^a$ shows how much overall performance changes by considering DMU i. If DMU i has no effect on the organization's overall performance, $\tilde{\theta}_i^a$ is equal to 1. If the overall performance increases (decreases) by considering DMU i, then $\tilde{\theta}_i^a > 1$ ($\tilde{\theta}_i^a < 1$). Since overall and partial performances can always be be computed, hyper-efficiencies cannot occur. However, in case of high values for $\tilde{\theta}$, super-efficiencies calculated by $\tilde{\theta}_i^a$ will not vary a lot. To derive more effective incentives, it might thus be reasonable to calculate super-efficiencies based on improvement potentials, given by the differences of the efficiency parameters to one ($\tilde{\theta}_i^b$). The ratios of improvement potentials fully correlate with $\tilde{\theta}_i^a$, but differ in variance. While the variance of $\tilde{\theta}_i^a$ decreases for increasing values of $\tilde{\theta}$, the variance of $\tilde{\theta}_i^b$ behaves vice versa. If super-efficiencies as $\tilde{\theta}_i^a$ or $\tilde{\theta}_i^b$ would be used to define optimal transfers according to (1), a branch would get an incentive to always act in the best interest of the entire organization and contribute as much as possible to overall performance.

3 Numerical Example

The applicability of our approach is demonstrated by evaluating the performances of 11 bank branches, using empirical data of a German bank from 2003. To compare

our results a basic DEA-model according to [2] is used. The on-balance liabilities from deposits and other bank accounts (**LIA**) as well as the employees' working hours of each branch (**EMP**) are considered as inputs. The Outputs are represented by all on-balance assets from credits (**ASS**) and the returns from non-credit businesses (**RET**). The original data, the efficiency and super-efficiency scores as well as the overall target values for inputs and outputs are summarized in table 1.

DMU	LIA	EMP	ASS	RET	θ_{VRS}	$\tilde{\theta}$	θ_{VRS}^{super}	$\tilde{\theta}^a$	$\tilde{\theta}^b$
1	14103	450	395	380	1,00	0,87	NaN	1,00	1,03
2	9158	218	3452	110	1,00		1,13	1,00	0,99
3	13819	530	8269	220	1,00		NaN	1,03	1,16
4	7469	130	1155	58	0,92		0,92	0,99	0,91
5	5646	150	756	75	0,95		0,95	0,98	0,89
6	5730	118	1110	85	1,00		1,16	1,00	1,02
7	8670	164	1406	89	0,80		0,80	0,99	0,96
8	5387	143	1201	102	1,00		1,17	1,01	1,06
9	10223	130	877	57	0,91		0,91	0,98	0,86
10	6041	130	1484	64	1,00		1,04	1,00	0,97
11	9464	130	855	48	0,91		0,91	0,97	0,83

Original and optimal production plans

	original	VRS	CRA
LIA	95710	83042	82842
EMP	2293	2219	1985
ASS	20960	21893	20960
RET	1288	1404	1288

Table 1 Centralized performance management: Original data, performances and targets.

θ_{VRS} denotes the efficiency in the BBC-model, θ_{VRS}^{super} the corresponding super-efficiencies. The BCC-model identifies 6 efficient branches (1,2,3,6,8,10) and 5 inefficient ones (4,5,7,9,11). Their average efficiency score is 0.95. CRA-DEA finds lager potentials for improvements and calculates an overall performance of 0.87. These differences in the input-oriented appraisal is reflected by the targets for the overall production plans, as shown on the right of table 1. It is noticeable that changes in the input as well as the output profile are suggested by the BCC-model, whereas outputs are kept constant and only inputs are changed by CRA-DEA. While the centralized model identifies a significantly lower consumption of inputs being feasible, the suggested output profile of the BCC-model is superior.[4]

In calculating super-efficiencies, the differences between the inter- and intra-organizational approach become even clearer. Firstly the occurrence of two cases of hyper-efficiency in the basic model is striking. The linear program is infeasible for the branches 1 and 3. For 7 of the remaining 9 branches the tendencies identified are the same in both models. However, for the branches 2 and 10 the results are ambiguous. While both are super-efficient in the BCC-model, their centralized super-efficiency scores are slightly below 1.[5] This suggests that with regard to overall

[4] However, in most industries, as in banking, output strongly depends on economic demand. Consequently, although an increase in outputs mathematically seems possible, it might not be achievable.

[5] Due to rounding, this is not displayed by $\tilde{\theta}^a$.

performance these branches pursue undesirable strategies. Both models also identify super-efficiencies at 4 branches, though not at the same. Only the branches 6 and 8 operate super-efficiently in both models, while their performance scores are higher in the ordinary DEA model.

4 Conclusion

DEA based Yardstick Competition has been argued to be of great use in deriving effective incentive mechanisms. However, due to methodical deficits of ordinary DEA models, the approaches cannot be transfered easily to internal performance management. Based on intra-organizational comparison procedures we developed two new approaches to calculate individual performance indicators and incentive mechanisms. Evaluating the performances of 11 bank branches, we have demonstrated how intra-organizational benchmarking can be applied.
However, intra-organizational performance management still is at its very beginning. Further efforts have to be made, for instance on exploring potential benefits and limits of integrating external information into a centralized performance evaluation. Also the applicability of CRA-DEA super-efficiencies to other problems remains to be investigated. Last but not least, the development of different incentive mechanisms based on other DEA models seem reasonable issues for further research.

References

1. P. Andersen and N.C. Petersen. A procedure for ranking efficient units in Data Envelopment Analysis. *Management Science*, 39: 1261–1264, 1993.
2. R. Banker, A. Charnes, and W. Cooper. Some Models for Estimating Technical and Scale Efficiency in Data Envelopment Analysis. *Management Science*, 30: 1078–1092, 1984.
3. P. Bogetoft. DEA-Based Yardstick Competition: The Optimality of Best Practice Regulation. *Annals of Operations Research*, 73: 277–298, 1997.
4. P. Bogetoft. DEA and Activity Planning under Asymmetric Information. *Journal of Productivity Analysis*, 13: 7–48, 2000.
5. L. Castelli, R. Pesenti, and W. Ukovich. A classification of DEA models when the internal structure of the Decision Making Units is considered. *Annals of Operations Research*, 173(1): 207–235, 2010.
6. S. Lozano and G. Villa. Centralized resource allocation using data envelopment analysis. *Journal of Productivity Analysis*, 22: 143–161, 2004.
7. A. Shleifer. A theory of Yardstick Competition. *The RAND Journal of Economics*, 16(3): 319–327, 1985.
8. J. Zhu. *"Quantitative Models for Performance Evaluation and Benchmarking"*. Springer, 2nd edition, 2009.

Non-Negativity of Information Value in Games, Symmetry and Invariance

Sigifredo Laengle and Fabián Flores-Bazán

Abstract In the context of optimization problems of an agent, *having more information without additional cost is always beneficial* is a classic result by D. Blackwell (1953). Nevertheless, in the context of strategic interaction this is not always true. Under what conditions more information is (socially) beneficial in games? Existing literature varies between two ends: on the one hand, we find works that calculate the information value of particular cases not always easy to generalize; on the other hand, there are also abstract studies which make difficult the analysis of more concrete applications. In order to fill this gap, we calculated the information value in the general case of constrained quadratic games in the framework of Hilbert spaces. As a result of our work we found a close relationship between the symmetry of information and the mathematical property of invariance (of subspaces). Such property is the base to calculate the information value and demonstrating its non-negativity. Such results improve the understanding of general conditions that assure the non-negativity of information value. In the immediate thing, this work can be extended to the study of other payoff functions with more general constraints.

1 Introduction

In the context of optimization problems of an agent, *having more information without additional cost is always beneficial* is a classic result by [5]. Nevertheless, in the context of strategic interaction this is not always true. Under what conditions more information is (socially) beneficial in games? How the characteristics of the players and the interaction among them affect the information value?

Sigifredo Laengle
Universidad de Chile, Diagonal Paraguay 257, Santiago de Chile e-mail: slaengle@fen.uchile.cl

Fabián Flores-Bazán
Universidad de Concepción, Casilla 160-C, Concepción, Chile e-mail: fflores@ing-mat.udec.cl

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_10,

Existing literature varies between two ends: on the one hand, we find works that calculate the information value of particular cases not always easy to generalize; on the other hand, there are also abstract studies which make difficult the analysis of more concrete applications. See for example [8, 3, 2] for the general Bayesian games case; and [4, 10] for a non-cooperative transportation network.

In order to fill this gap, we calculated the information value in the general case of constrained quadratic games in the framework of Hilbert spaces; we determined the conditions to assure its no-negativity. Many classical situations are subsumed by our general model. The Cournot duopoly game [12] and its extension of *duopoly with differentiated products* [6, 11] are examples of applications. Pursuit evasion [9] can be modeled as an approximation game similar to the one considered in the present paper, as well as in Hinich and Enelow's spatial voting theory [7], where the authors model voters' objective function as an approximation of their ideal policy.

As a result of our work we found a close relationship between the symmetry of information and the mathematical property of invariance (of subspaces). Such property is the base to calculate the information value and demonstrating its non-negativity. Such results improve the understanding of general conditions that assure the non-negativity of information value. In the immediate thing, this work can be extended to the study of other payoff functions with more general constraints.

2 Non-Negativity of Information Value

To obtain useful and plausible conclusions, we pose and solve a general quadratic model with linear constraints in Hilbert spaces. On the one hand, the power of abstract models in Hilbert spaces allows us to consider a wide range of stochastic and dynamic problems. On the other hand, the linear-quadratic formulation allows the calculation of the value of information and specify conditions to ensure non-negativity without losing a wide range of applications. In this section we propose the model, define the social value of information, discuss the condition of invariance as the main result and introduce the interesting case of a duopoly with differentiated products [6, 11] as an application.

The Model. Let V_i be the **event space** for player i ($i = 1,2$), which is Hilbert, both endowed with the inner product denoted by the same symbol $\langle \cdot, \cdot \rangle$. Let R be the **resources space** as Hilbert space too. For each $i = 1,2$, let $A_i : V_i \to R$ be the **technology operator** of player i, being bounded and linear that transforms the strategies into resources. Similarly, for $i \neq j$, let the **team operator** $M_i : V_i \to V_i$, which is bounded and linear. Furthermore, we define the (i,j)-**interaction operator** between both players by $N_i : V_j \to V_i$. The element $u = (u_1, u_2) \in V_1 \times V_2$, where u_i is the individual **objective** of player i, is given. Let $E_1 \subseteq V_1$, $E_2 \subseteq V_2$, be the **strategies spaces** of player 1 and 2, which are closed subspaces; the **available resource** $b \in \mathcal{R}(A) = A(V_1 \times V_2)$ is given, where $A : V_1 \times V_2 \to R$ is defined by $A(x_1, x_2) = A_1 x_1 + A_2 x_2$.

Set $V \doteq V_1 \times V_2$. This is a Hilbert space too endowed with the inner product $\langle v,w\rangle \doteq \langle v_1,w_1\rangle + \langle v_2,w_2\rangle$ if $v_1 = (v_1,v_2) \in V_1 \times V_2$, $w = (w_1,w_2) \in V_1 \times V_2$. The bounded linear **team operator** $M : V \to V$ defined by $M(v_1,v_2) = (M_1v_1, M_2v_2)$ and the bounded linear **interaction operator** $N(v_1,v_2) = (N_1v_2, N_2v_1)$. Let us denote by $E \doteq E_1 \times E_2$ and $K_E \doteq \{x \in E : Ax = b\}$. By assumption $K_E \neq \emptyset$. The case $A = 0$, $b = 0$ corresponds when no constrain on the strategies is imposed.

The Nash equilibrium problem on E, $Q_2(E)$, consists in finding $\bar{x} = (\bar{x}_1,\bar{x}_2) \in K_E$ such that

$$\left.\begin{array}{l} f_1(\bar{x}_1,\bar{x}_2,u_1) \leq f_1(x_1,\bar{x}_2,u_1) \text{ for all } x_1 \in E_1,\ A_1x_1 + A_2\bar{x}_2 = b \\ f_2(\bar{x}_1,\bar{x}_2,u_2) \leq f_2(\bar{x}_1,x_2,u_2) \text{ for all } x_2 \in E_2,\ A_1\bar{x}_1 + A_2x_2 = b \end{array}\right\} \qquad (Q_2(E))$$

Such an $\bar{x}$ is called a Nash equilibrium solution. The remainder of this section considers the following **loss functions**

$$f_i(x_1,x_2,u_i) \doteq \frac{1}{2}\langle M_ix_i,x_i\rangle - \langle u_i - N_ix_j,x_i\rangle \ (i \neq j).$$

Obviously, the function f_i is convex in x_i. For a given $(\bar{x}_1,\bar{x}_2) \in K_E$, we set $K_{E_1} \doteq \{x_1 \in E_1 :\ A_1x_1 + A_2\bar{x}_2 = b\}$, $K_{E_2} \doteq \{x_2 \in E_2 : A_1\bar{x}_1 + A_2x_2 = b\}$, which are convex and closed sets. Obviously $K_{E_1} \times K_{E_2} \subseteq K_E$.

An interesting case we have developed in this framework is a bayesian game as follows. Let us consider two probability spaces, one for each player. For $i = 1,2$, let us suppose Ω_i to be a (non empty) set of states of the nature and a structure of information $\mathcal{B}_i$ defined as a σ-algebra of Ω_i. We define the probability spaces $(\Omega_i,\mathcal{B}_i,\mathbf{P}_i)$ and the events space V_i as $\mathcal{L}^2(\Omega_i,\mathcal{B}_i,\mathbf{P}_i)$, that is, the set of $\mathcal{B}_i$-measurable random variables defined in Ω_i with finite variance. Let us consider following sub-σ-algebras of $\mathcal{B}_i$, $\mathcal{E}_i$ and $\mathcal{F}_i$, such that $\mathcal{E}_i \subseteq \mathcal{F}_i \subseteq \mathcal{B}_i$. $\mathcal{E}_i$ can be interpreted as the information structures less informative, and $\mathcal{F}_i$ more informative. The strategies spaces E_i are given by $\mathcal{L}^2(\Omega_i,\mathcal{E}_i,\mathbf{P}_i)$, a subspace of $\mathcal{L}^2(\Omega_i,\mathcal{B}_i,\mathbf{P}_i)$ in the case less informative and F_i by $\mathcal{L}^2(\Omega_i,\mathcal{F}_i,\mathbf{P}_i)$ in the case more informative. The loss function of player i is expressed as

$$\mathbf{E}(\frac{1}{2}m_ix_i^2 - (u_i - n_ix_j)x_i), \quad \text{where } i \neq j \text{ and } u_i \in \mathcal{L}^2(\Omega_i,\mathcal{B}_i,\mathbf{P}_i).$$

Further, we consider that players are faced to a common constrain of resource, given by $\mathbf{E}(a_1x_1 + a_2x_2) = b$, i.e., the expected value of the use of resource of both players is limited to a available quantity $b \in \mathbb{R}$. Let us point out that the parameters m_i (of the interaction operator), n_i (of the interaction operator), and a_i (of the technology matrix) will play an important role in our discussion below. Let us observe that, in this example, the technology matrix $A : V_1 \times V_2 \to \mathbb{R}$ is given by $A(v_1,v_2) \doteq \langle a_1,v_1\rangle + \langle a_2,v_2\rangle = \mathbf{E}(a_1v_1) + \mathbf{E}(a_2v_2) = a_1\mathbf{E}(v_1) + a_2\mathbf{E}(v_2)$, where $a_i \in \mathbb{R}$. Hence the adjoint $A^* : \mathbb{R} \to V_1 \times V_2$ is given by $A^*b = b(a_1,a_2)$. Furthermore, the team M and the interaction operators are given by $M(v_1,v_2) = (m_1v_1, m_2v_2)$ and $N(v_1,v_2) = (N_2v_1, N_1v_2)$ respectively, where m_i and n_i $(i \neq j)$ are a real numbers.

Defining and Computing the Social Information Value. We now define the (social) information value for the problem $Q_2(E)$. Let F_1, F_2 be two strategies subspaces of the event space V such that $E_1 \subseteq F_1$ and $E_2 \subseteq F_2$. This is interpreted as the more informed situation (F_i) and the less informed (E_i). Let $\bar{x} = (\bar{x}_1, \bar{x}_2) \in K_E$, $\bar{y} = (\bar{y}_1, \bar{y}_2) \in K_F$ (with $F = F_1 \times F_2$) be Nash equilibrium solutions to problem $Q_2(E)$ and $Q_2(F)$ respectively. We define the **(social) information value of problem Q_2 of F with respect to E** by

$$I_2(E,F) \doteq \underbrace{f_1(\bar{x}_1,\bar{x}_2) + f_2(\bar{x}_1,\bar{x}_2)}_{\text{Solution in } E} - \underbrace{(f_1(\bar{y}_1,\bar{y}_2) + f_2(\bar{y}_1,\bar{y}_2))}_{\text{Solution in } F}.$$

Now, our aim is to establish an explicit formula for the information value. To that end, we need to introduce of the notion of observability of the subspaces of strategies and of resources.

Given $E_1 \subseteq V_1$ and $E_2 \subseteq V_2$ spaces of strategies. Let M_i be the team operator of player i. We say that the strategies subspace E_i is **observable** through the operator M_i iff $M_i(E_i) \subseteq E_i$, which corresponds to the definition of *invariance* of closed subspace of a Hilbert space by an linear bounded operator. A similar definition is given for the interaction operator N_i. It is not difficult to verify that the observability of the strategies subspace $E \doteq E_1 \times E_2$ through $M+N$ is equivalent to the observability of E_i through M_i, i.e. $M_i(E_i \subseteq E_i)$, and the observability of E_j through N_i, i.e. $N_i(E_j) \subseteq E_i$.

Similarly, let $A_i : V_i \to R$ $(i = 1,2)$ the technology operator, we say that the resource space R is **observable** to the players through (A_1^*, A_2^*) if $(\mathcal{N}(A_1) \times \mathcal{N}(A_2))^\perp \subseteq E_1 \times E_2$. It is equivalent to say that R is observable to each player i, i.e., if $\mathcal{N}(A_i)^\perp \subseteq E_i$ for $i = 1,2$, since $\mathcal{N}(A_1)^\perp \times \mathcal{N}(A_2)^\perp = (\mathcal{N}(A_1) \times \mathcal{N}(A_2))^\perp$. Due to the inclusion $\mathcal{N}(A_1) \times \mathcal{N}(A_2) \subseteq \mathcal{N}(A)$, where $A(x_1,x_2) \doteq A_1(x_1) + A_2(x_2)$, we conclude that, if R is observable through (A_1^*, A_2^*) then R is observable through A^*, i.e., $\mathcal{N}(A)^\perp \subseteq E_1 \times E_2$. Furthermore, it is not difficult to check that $A^* : R \to V_1 \times V_2$ is given by $A^*q = (A_1^*q, A_2^*q)$, $q \in R$.

Furthermore, let us introduce orthogonal projectors P_{E_i} from V_i onto closed subspaces E_i respectively. Each P_{E_i} is a linear operator, idempotent ($P_{E_i} \circ P_{E_i} = P_{E_i}$), and self-adjoint ($\langle P_{E_i} v_i, w_i \rangle_i = \langle v_i, P_{E_i} w_i \rangle_i$ for all $v_i, w_i \in V_i$). In the following theorem, we use the orthogonal projector P_E defined from $V \doteq V_i \times V_i$ onto $E_i \times E_i$ by $P_E(v_1,v_2) \doteq (P_{E_1}v_1, P_{E_2}v_2)$.

Now, let us establish the main theorem of this section, namely the formula for computing the information value under observability assumptions.

Theorem 1 *Let, for $i = 1,2$, $E_i \subseteq V_i$, $F_i \subseteq V_i$ be strategies subspaces. Set $E = E_1 \times E_2$, $F = F_1 \times F_2$, with $E \subseteq F$, and assume that: (1) R is observable through (A_1^*, A_2^*); (2) $E, F, \mathcal{N}(A)$ are observable through $M+N$ and $(M+N)(\mathcal{N}(A)) \subseteq \mathcal{N}(A_1) \times \mathcal{N}(A_2)$. If $\bar{x}$ and $\bar{y}$ are Nash equilibrium solutions to $Q_2(E)$ and $Q_2(F)$ respectively, then the social information value is given by*

$$I_2(E,F) = \frac{1}{2}\langle M^{-1}(P_F - P_E)u, (P_F - P_E)u\rangle - \langle (M+N)^{-1}(P_F - P_E)u, (P_F - P_E)u\rangle + \langle M(M+N)^{-1}(P_F - P_E)u, (M+N)^{-1}(P_F - P_E)u\rangle$$

which is non-negative, i.e. $I_2(E,F) \geq 0$.

The proof of the non-negativity of the formula of $I_2(E,F)$ is based on the basic assumption of observability and requires intermediate results. This assumption has a lot of implications, which are not developed here. For example, the Nash solution in the *more informed* case (F) can be obtained from the Nash solution in the *less informed* case (E) by the formula $(M+N)\bar{y} = (M+N)\bar{x} + P_F u - P_E u$, which is an interesting result by itself.

Discussion of the Main Assumptions of Theorem 1. The first condition of Theorem 1 for ensuring the non-negativity of information value considers the capacity of the game to observe the resources space through the adjoint of its technology matrix, i.e., $A^*(R) \subseteq E$, where E is the strategies subspace. This *capacity of the game* is equivalent to the capacity of each player to observe individually the common resources space R, because $A^*(R) \subseteq E$ is equivalent to $A_i^*(R) \subseteq E_i$. This condition can be interpreted as follows: *if the technology matrix is determined by observable elements, then the observability of resources is satisfied.*

The second condition of Theorem 1 for ensuring the non-negativity of information value consists in the observability of the subspaces E, F and $\mathcal{N}(A)$ through the interaction operator $M+N$. Similarly to the first condition, this one means that *the team and interaction operators do not affect what each player observes*. Furthermore, this observability condition of E under M is equivalent by requiring that $M_i(E_i) \subseteq E_i$ and $N_i(E_j) \subseteq E_i$, $i \neq j$, meaning that *each player i is able to observe her strategies spaces through the team operator M_i and the strategies of the other player through its interaction operator N_i.*

Duopoly with Differentiated Products. Following [6, 11], consider a two-firm industry producing two differentiated products indexed by i ($i = 1,2$). To simplify the exposition, we assume that production is costless. We assume the following (inverse) demand for product i ($i = 1,2$) is given by the prices

$$u - m_i v_i - n_i v_j \; m_i > 0, m_i > n_i, \; i \neq j,$$

where $v_i \geq 0$ is the production level for each firm. Thus, we assume that each product is produced by a different firm facing the given demands. The assumption $m_i > n_i$ means that the effect of increasing v_i on the price of product i is larger than the effect of the same increase on the price of product j. This is described by saying that *own-price effect* dominates the *cross-price effect*. We note that each firm i choose the production level to maximize their utility. In this example the loss functions are given by

$$f_i(v_i, v_j) \doteq v_i(m_i v_i + n_i v_j - u) \quad i,j = 1,2; \; i \neq j,$$

and the equilibrium $(\overline{v}_1, \overline{v}_2)$ satisfies

$$f_i(\overline{v}_i, \overline{v}_j) = \min\{v_i(m_i v_i + n_i \overline{v}_j - u) : v_i \geq 0\} \quad i, j = 1,2;\ i \neq j.$$

Thus the scheme is described by our theory.

Classic Cournot's Duopoly. An interesting example is the classic Cournot's duopoly, which can be easily deduced from the duopoly's example with differentiated products. Indeed, if $m_i = n_j = 1$ for i, j $(i \neq j)$, we obtain the loss functions of classic duopoly as described on p. 108 of [1]:

$$f_1(v_1, v_2) = v_1(v_1 + v_2 - u) \quad \text{and} \quad f_2(v_1, v_2) = v_2(v_1 + v_2 - u).$$

Acknowledgements The first author, was partially supported by the Department of Mathematical Engineering, Universidad de Concepción (Chile), during 2009, and was funded by Department of Information System, Universidad de Chile. The author thanks the Department of Mathematical Engineering for its hospitality.

References

1. J. Aubin. *Optima and equilibria: an introduction to nonlinear analysis*. Springer, 1998.
2. B. Bassan, O. Gossner, M. Scarsini, and S. Zamir. Positive value of information in games. *International Journal of Game Theory*, 32(1): 17–31, 2003.
3. B. Bassan, O. Gossner, M. Scarsini, and S. Zamir. A class of games with positive value of information. *THEMA Working Papers*, 1999.
4. N. Bean, F. Kelly, and P. Taylor. Braess's paradox in a loss network. *Journal of Applied Probability*, 34(1): 155–159, 1997.
5. D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2): 265–272, 1953.
6. A. Dixit. A model of duopoly suggesting a theory of entry barriers. *The Bell Journal of Economics*, 10(1): 20–32, 1979.
7. J. M. Enelow and M. J. Hinich. *The spatial theory of voting: an introduction*. Cambridge University Press, New York, 1984.
8. O. Gossner. Comparison of Information Structures. *Games and Economic Behavior*, 30(1): 44–63, 2000.
9. R. Isaacs. Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization. Dover Publications, New York, 1999.
10. Y. A. Korilis, A. A. Lazar, and A. Orda. Avoiding the Braess paradox in non-cooperative networks. *Journal of Applied Probability*, 36: 211–222, 1999.
11. N. Singh and X. Vives. Price and quantity competition in a differentiated duopoly. *The RAND Journal of Economics*, 15(4): 546–554, 1984.
12. Jean Tirole. *The theory of industrial organization*. The MIT Press, New York, 1988.

Smart Entry Strategies for Markets with Switching Costs

Florian W. Bartholomae, Karl Morasch, and Rita Orsolya Toth

Abstract In this paper we consider a market with switching costs that is initially served by a monopolistic incumbent. How can a competitor successfully enter this market? We show that an offer to undercut the incumbent by a fixed margin serves this purpose. This strategy dominates traditional entry where the entrant just offers a lower price because it restrains the ability of the incumbent to block entry by limit pricing. We also consider adding a price ceiling to insure customers against future price increases. This strategy turns out to be the preferable one for entering markets with elastic demand.

1 Basic Model: Players, Strategies and Timing

In the basic model, we consider a three–stage game as depicted in fig. 1. There are three players: a monopolistic incumbent (I), an entrant (E) and consumers, who are initially served by the incumbent. Both firms are retailers of a homogeneous product x with inverse demand given by $p(x)$. Hence, price competition between the two firms would result in the well-known Bertrand-paradox, if there are no switching costs and procurement costs of both sellers are identical. However, the assumption that consumers have to bear switching costs d results in an asymmetry between entrant and incumbent.

With the exception of the second stage under price competition this is a game of perfect information, which therefore may be solved by backward induction. In the first stage the entrant has to choose between three different entry strategies:

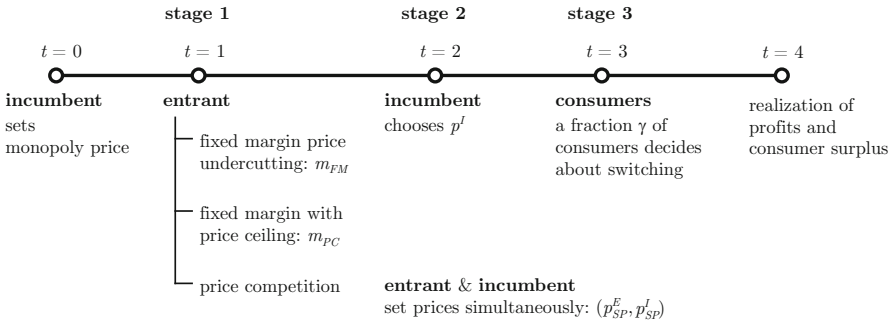
Florian W. Bartholomae
Universität der Bundeswehr München, e-mail: florian.bartholomae@unibw.de

Karl Morasch
Universität der Bundeswehr München, e-mail: karl.morasch@unibw.de

Rita Orsolya Toth
Universität der Bundeswehr München, e-mail: rita.toth@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_11,

Fig. 1 Time structure of the model: Players and strategies



- **Strategy FM** The first option is to undercut the incumbent's price by a fixed margin. This margin will be maintained if the incumbent changes its price in the second stage, resulting in a price $p^E_{FM}(p^I) = p^I - m_{FM}$.
- **Strategy PC** The entrant may supplement the margin by a price ceiling that is determined by the monopoly price p^M and the chosen undercutting margin. This implies that the entrant cannot increase its price if the incumbent chooses a price above the initial monopoly price (either due to rising procurement costs or strategic considerations). However he is obliged to lower his price to maintain the initially set margin whenever the incumbent's price decreases. The resulting price is thus given by $p^E_{PC}(p^I, p^M) = \min\left\{p^I - m_{PC}, p^M - m_{PC}\right\}$.
- **Strategy SP** Finally, the new firm can enter the market as a "normal" competitor. In this case he will compete with the incumbent in a simultaneous pricing game in the second stage that yields prices (p^I_C, p^E_C).

Note, that the strategies differ with respect to commitment ability and timing as indicated in fig. 1. We assume that the entrant is able to credibly commit to the chosen margin under strategies FM and PC. An announcement to stick to the margin will be binding if it is made in a way that is enforceable by a court. There is, however, a second commitment issue: If the incumbent lowers his price below the margin, the entrant would incur a loss that could force him into bankruptcy. We therefore assume that the entrant is known to have enough funds to incur losses up to the amount of the margin times the number of units sold when serving all customers. Given this, it is not possible for the incumbent to drive the entrant out of the market without incurring a loss for himself. As the incumbent may react in stage 2 to any price offer by a "normal" entrant (strategy SP) by lowering his own price, it seems to be most sensible to model this situation as simultaneous price setting game. Here neither the incumbent nor the entrant can commit themselves to a certain price. In the case of the other two strategies the fixed margin proposed by an entrant ensures that his price is automatically adjusted whenever the incumbent decides to alter his own price. This yields a sequential structure where the entrant sets his margin first and the incumbent optimally reacts to the given margin. The optimal strategy for the

entrant is then given by the critical margin $m^*(p^I)$ that is just necessary to induce switching.

For all entry strategies we assume in stage 3 that a fraction $\gamma \in (0,1]$ of all consumers observes the resulting price offers. These consumers decide whether they switch to the entrant or stay with the incumbent. Note that a value of $\gamma < 1$ seems to be sensible description of reality as due to search costs not all consumers will frequently search for competing offers in such markets. The prices that result from the behavior of entrant and incumbent in stages 1 and 2 determine the realized individual demand for the strategies "switch to entrant" and "remain with the incumbent". It is then straightforward to determine the payoffs based on prices and realized demand.

2 Switching under Inelastic Demand

In a first step we want to highlight the working of entry by fixed margin price undercutting and compare it to the entry with strategy SP in a stripped–down model. For simplicity we assume inelastic demand and identical consumers with a maximum valuation v. In the period considered, each consumer is assumed to buy exactly one unit of the good as long as the price does not exceed her net valuation. Here, the minimal margin to induce switching just equals d. Furthermore, at first we consider a situation where $\gamma = 1$, i. e. all consumers consider switching.

Without competition the monopolistic incumbent maximizes his profit by setting the price to $p^M = v$. Note, that in this simple setting the monopolist obtains the complete rent and the market is efficiently served. Therefore, any switching by consumers to the entrant will yield a welfare loss due to switching costs. Under the "traditional" entry strategy we get the well–known results for price competition with asymmetric costs.

Proposition 1 (No switching under price competition)
Price competition with switching costs yields prices $p_C^I = d$ and $p_C^E = 0$. All consumers stay with the incumbent.

Proof. If the monopolist charges a price $p^I \in [p^M, d)$, the entrant would set a price $p^E = p^I - d - \varepsilon$ with $\varepsilon \to 0$. As a result, all consumers would switch to the entrant and the profit of the incumbent would equal zero. The entrant's incentive to charge a lower price can only be avoided if the incumbent lowers his price to $p^I = d$. In order to induce switching, the entrant would then need to reduce his price to $p^E = -\varepsilon$. However, this is not optimal as a price below zero would yield a negative profit. □

There is no entry in this setting, only a reduction in prices due to potential competition as the incumbent blocks entry by applying a limit pricing strategy.

We obtain a completely different result if we assume entry by fixed margin price undercutting. In this case the entrant sets his margin m first, then the incumbent reacts by setting his price p^I, and finally consumers decide about switching based

on prices p^I and p^E_{FM}. While the incumbent might still be able to lower his price far enough to induce negative profits for the entrant, the entrant is committed and cannot be driven out of the market. This yields a quite strong result.

Proposition 2 (Fixed margin price undercutting is effective)
The entrant sets m marginally greater than d. The incumbent stays at price p^M and all consumers switch to the entrant. The entrant earns a surplus $p^M - m$ per consumer.

Proof. In order to induce consumers to switch, the margin m has to be greater than the switching costs d. Therefore the entrant will set his margin at $m = d + \varepsilon$, with ε arbitrarily close to zero. If the monopolist faces this strategy, he cannot improve his situation by changing the price relative to the pre–entry value p^M. Reducing his price would just lower the entrant's price by the same amount. As the margin m stays constant and exceeds the switching costs, all consumers are going to switch to the entrant. □

While entering the market in the traditional way is not a successful strategy, fixed margin price undercutting not only allows the entrant to enter the market, but he actually replaces the incumbent. As will be shown, strategies FM and PC dominate SP in almost all settings as an entry strategy. However, results are much less extreme when we no longer assume complete switching and inelastic, individual demand.

Let us first consider limited switching. In the given setting either all consumers stay with the incumbent or all switch to the entrant. In reality, only a fraction of the consumers is likely to switch, which may be either due to differences in switching costs or in the consumers' awareness of a competing supplier. Hence, we assume that at a given point of time only a fraction of all consumers contemplate about changing their supplier. In this case it may no longer be in the interest of the incumbent to deter entry by a limit price as this would reduce his profits from the consumers that will stay with him anyway. As shown in [1] entry with strategy SP will thus take place, if $d > (1 - \gamma)v$.

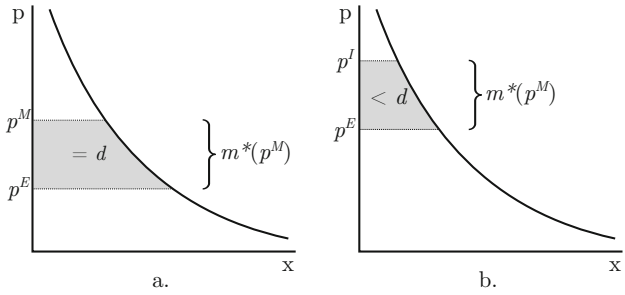
The attempt of this section was to highlight the difference between fixed margin price undercutting and traditional price competition under the assumption of inelastic demand and both complete and limited switching by consumers. We will now generalize our analysis by considering a specification with elastic demand.

3 Elastic Demand, Limited Switching and Price Ceiling

The case with elastic individual demand extends the incumbent's strategical options compared to the previous situation. Here, the incumbent would have an incentive to prevent switching by slightly raising his price in stage 2, if the entrant has chosen his margin based on the initial monopoly price. This situation is illustrated in figure 2. As figure 2a. shows, the minimal margin that is sufficient under the monopoly price, p^M, to compensate for the switching cost no longer assures a sufficient increase of

consumer surplus under the higher price, p^I. Hence, under p^I the minimal margin will no longer compensate a consumer for her individual switching costs (figure 2b). In this setting we ask two questions: Will the monopolist be able to block entry with fixed margin price undercutting? Can a price ceiling serve as a means against an entry blocking strategy?

Fig. 2 Individual demand curves: Increasing the monopoly price raises the minimal margin



As before we determine the subgame perfect Nash–equilibria by backward induction. While the decision problem of a consumer in stage 3 is still straightforward, the analysis of the price setting behavior of the incumbent in stage 2 becomes more intricate. Contrary to the case of inelastic demand, the incumbent now has the possibility to react to the margin set by the entrant, and in turn to deter consumers from switching. However, unlike in the situation under price competition, the incumbent does not reduce his price to prevent entry in the case of strategy FM. Instead, he has an incentive to raise the price to reduce the advantage that a consumer may get from switching for the given margin: lowering the price would not help, as the entrant is not committed to the price but to his margin, and a given margin will result in a higher change in consumer surplus if p^I is reduced.

While the potential reactions of the incumbent go in different directions for fixed margin price undercutting and entry with traditional price competition, respectively, the impact on the price of a successful entrant is similar. To prevent any of the two forms of limit pricing, the entrant must charge a lower price than $p^M - d$ (either through directly reducing the own price or by choosing a higher margin). As it can be shown (for details see [1]), the incumbent will not increase his price in equilibrium as the entrant will raise the margin sufficiently to make this strategy unprofitable. Note, however, that the entrant can do better by introducing a price ceiling that is based on the initial monopoly price. In this case, he can just set the minimal margin for p^M as the actual margin would increase automatically, if the incumbent raises its price above p^M.

Proposition 3 (Dominance of a price ceiling)
In a setting with a downward sloping demand curve the incumbent has an incentive to prevent market entry by increasing his price in $t = 2$. The entrant can avoid this

price increase either through setting a higher margin or by choosing a price ceiling. In both cases the equilibrium strategy of the incumbent is to charge the monopoly price. However, as the PC strategy entails a lower margin, it dominates fixed margin price undercutting without a price ceiling.

Proof. Under the FM strategy the entrant has to account for the incumbent's reaction when choosing the margin. With downward sloping demand the incumbent could reduce the change in consumer surplus for a given margin by raising his price. Hence, the optimal margin must be based on a price that is high enough to make the incumbent be indifferent between selling to all consumers at this price and charging the monopoly price to the remaining $(1-\gamma)$ customers who do not consider switching. This is not necessary with a price ceiling where the margin could be set according to the monopoly price. Since the chosen level of the margin is higher under "pure" fixed margin price undercutting than under the PC strategy, the corresponding profits under strategy FM are lower. If switching costs are low enough to ensure positive profits for the entrant under both strategies, the consumers that consider switching actually switch to the entrant in equilibrium under both strategies, as the incumbent has either no incentive (FM strategy) or no possibility (PC strategy) to prevent them from doing so. □

In order to compare profits under the discussed strategies, we did some numerical simulations (see [1]). According to our results, for relatively low switching costs, successful entry into the market is ensured for all three strategies. Above a certain threshold for switching costs, the incumbent chooses the limit-pricing strategy under entry strategy SP. The traditional entry strategy results generally in substantially lower profits than strategies FM and PC. In a situation without switching costs strategy FM already yields the highest feasible profit, which implies that there is no necessity for a price ceiling. However, with rising switching costs the advantage of the strategy with price ceiling becomes more and more pronounced.

4 Conclusion

In this paper we have shown that not only incumbents but also entrants have the opportunity to improve their competitive position by a strategic move. Binding the own price to the incumbent's price by a fixed margin may be a smart strategy that restricts the options of an incumbent in a way that he is no longer able to deter entry.

References

1. Florian W. Bartholomae, Karl Morasch, and Rita Orsolya Toth. *Smart Entry in Local Retail Markets for Electricity and Natural Gas*, volume 21–3 of *Universität der Bundeswehr München Working Papers in Economics*. 2009.

A Fresh Look on the Battle of Sexes Paradigm

Rudolf Avenhaus and Thomas Krieger

Abstract If one tries to model real conflict situations with the help of non-cooperative normal form games, it may happen that strategy combinations have to be considered which are totally unrealistic in practice but which, however may be taken into account in equilibrium with positive probability.
In this paper the battle of sexes paradigm is considered which is the most simple game owning this unrealistic feature. It is shown that a slight modification of the rules of this game remedies the problem: The game is repeated as long as the absurd strategy combination is realized, but at most n times. It turns out that the expected run length of this new game is smaller than two. In other words, the unrealistic feature of the original battle of sexes paradigm can be removed by a slight and in its consequences not important modification of the rules of the game.

1 Introduction

When we tried to describe the conflict between the North and the South of the Sudan with the help of a non-cooperative normal form game, strategy combinations had to be considered which appeared to be totally unrealistic in practice, see [6]. Therefore, we gave up this approach but continued to find out in which way conflict situations were modelled where similar problems occurred. Indeed, in the theory of correlated equilibria situations of this kind are discussed, but there, the existence of a mediator or a noisy channel is required, see, e.g., [7], which does not hold for situations like the Sudan conflict. Thus, we looked for an intrinsic solution.

Rudolf Avenhaus
Universität der Bundeswehr München, Fakultät für Informatik, Werner-Heisenberg Weg 39, 85579 Neubiberg, Germany, e-mail: rudolf.avenhaus@unibw.de

Thomas Krieger
ITIS GmbH an der Universität der Bundeswehr München, Fakultät für Informatik, Werner-Heisenberg Weg 39, 85579 Neubiberg, Germany e-mail: thomas.krieger@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_12,

For the sake of the argument we consider in the following the most simple conflict situation with this difficulty, namely the well-known battle of sexes paradigm.

For this non-cooperative two-person game there exist two equilibria in pure strategies, and one in mixed strategies, which are realized by an appropriate random experiment. If both players agree on the latter one, which provides the same payoff to both players, then with positive probability the absurd situation occurs that both players choose independently of each other that strategy they like the least.

In the following a modification of this game is proposed such that the random experiment is repeated as long as the absurd strategy combination is realized, but at most n times (e.g., in order to limit quarrelling time). It will be shown that the expected number of repetitions is smaller than two for $n \geq 2$. In other words, the unrealistic feature of the original Battle of Sexes paradigm can be removed by a slight and in its consequences not important modification of the rules of the game.

We conclude this paper with some remarks about the applicability of these results to more realistic and complicated conflict situations with the above described property.

2 Original Model

Assume that a couple cannot agree how to spend the evening together, see [3] and also [5]: He wants to attend a boxing match whereas she wants to go to a ballet. Of course, both would like to spend the evening together. In Figure 1 the normal form of this non-cooperative two-person game is shown.

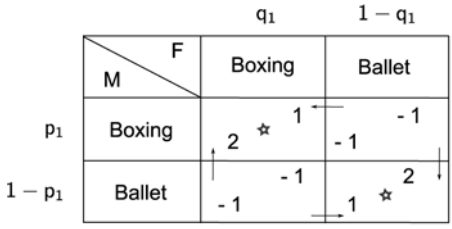


Fig. 1 Normal form of the original Battle of Sexes paradigm. The arrows indicate the preference directions, the stars denote the two equilibria in pure strategies.

As shown in Figure 1, there are two Nash equilibria in pure strategies and one in mixed strategies: Let M_1 and F_1 be the expected payoffs to both players. Then the mixed Nash equilibrium is given by

$$p_1^* = \frac{3}{5}, \quad q_1^* = \frac{2}{5}, \quad M_1^* = F_1^* = \frac{1}{5}. \tag{1}$$

Two problems characterize this model: First, there are three Nash equilibria none of which can be chosen in a natural way. This should, however, not be considered a weakness of the model, but a representation of reality: Otherwise there would be no quarrel. The bargaining model by Nash (1951) provides one unique solution, but is totally different.

Second, in case the mixed equilibrium is agreed upon – both players get the same payoffs – with the positive probability $q_1^*(1-p_1^*) = (2/5)^2 = 0.16$ the absurd situation occurs that the man attends the ballet and the wife the boxing match. This is quite unrealistic therefore, in the following we will consider this problem.

Nevertheless, such absurd situations occur in reality: G. O. Faure, Sorbonne University, attended a meeting of French and German cultural delegations in Paris, see [1]. On the morning of the meeting which was scheduled for 9 am, the French delegation entered the meeting room at 9 sharp and the German delegation at 9:15. The next day the same thing happened again. Obviously, both delegations wanted to show respect for each other. The French, assuming the Germans always arrive in time did not want to let them wait. Conversely the Germany assuming that French are late, did not want to blame them by arriving early. We will return to this case in the fourth section.

3 New Model

Let us assume that the couple agrees to choose the symmetric, i.e., the mixed Nash equilibrium and furthermore, to repeat the game in case an appropriate random generator realizes the absurd strategy combination. One could object that then both better would agree on a random experiment which results in a joint visit either of the boxing match or the ballet, but this might go too far since it would exclude the separate visit of both events. Also let us mention that we tried, so far without success, to formulate some axiomatic foundations of the rule of the new game.

To begin we consider a game which consists of only two steps. If we represent this new game in extensive form, then we see that the second step game is a subgame of the total one which means that the equilibrium can be determined recursively with the help of backward induction, see [5].

Let us continue this way: We define the n-step game such that i) in case the absurd strategy combination is realized by the random generator on steps 1 to $n-1$, the random experiment is repeated and ii) the game ends the latest with the n-th step which may mean, that then the absurd strategy combination is realized and accepted as solution of the game. Let p_{n-i+1} resp. q_{n-i+1} be the probability that the man resp. the woman chooses boxing on the i-th step, $i = 1,\ldots,n$, and let M_n resp. F_n be the expected payoff of the man resp. the woman. The reduced normal form of this game is shown in Figure 2.

Lemma 1 *The equilibrium strategy $(p_n^*,\ldots,p_1^*)$ resp. $(q_n^*,\ldots,q_1^*)$ and the corresponding payoff M_n^* resp. W_n^* of the man resp. the woman are given by the recursive relations*

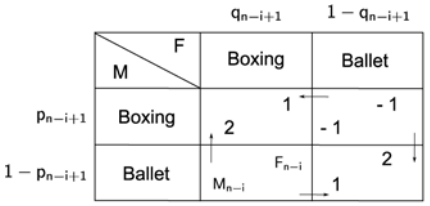


Fig. 2 Reduced normal form of the i-the step of the n-step Battle of Sexes paradigm.

$$p_j^* = \frac{2-F_{j-1}^*}{4-F_{j-1}^*}, \quad F_j^* = \frac{2+F_{j-1}^*}{4-F_{j-1}^*}, \quad j=1,\ldots,n, \quad F_0^* = -1 \tag{2}$$

$$q_j^* = \frac{2}{4-M_{j-1}^*}, \quad M_j^* = \frac{2+M_{j-1}^*}{4-M_{j-1}^*}, \quad j=1,\ldots,n, \quad M_0^* = -1\,. \tag{3}$$

Proof. The proof is given in [2]. □

The explicit solution of the recursive relations (2) and (3) is given by the following

Lemma 2 *The mixed equilibrium payoffs to the two players of the n-step game are*

$$M_n^* = F_n^* = 2 - \frac{1}{1-(2/3)^{n+1}}\,.$$

Proof. This relation is proven by verifying (2) resp. (3). □

Let for the n-step game w_i^*, $i=1,\ldots,n$ be defined as the probability that on the $(n-i+1)$-step the absurd strategy is realized (steps are counted backward!)

$$w_i^* = q_i^* \cdot (1-p_i^*) = \left(\frac{2}{4-M_{i-1}^*}\right)^2, \quad i=1,\ldots,n\,. \tag{4}$$

The probability, that in the n-step game the run length $L^* = l$, $l=1,\ldots,n$, is

$$P(L_n^* = l) = \begin{cases} 1-w_n^* & : \quad l=1 \\ \prod_{j=n-l+2}^{n} w_j^* \cdot (1-w_{n-l+1}^*) & : \quad l=2,\ldots,n-1 \\ \prod_{j=2}^{n} w_j^* & : \quad l=n \end{cases}\,. \tag{5}$$

According to (3), (4) and (5) the expected run length $E(L_n)$ of the n-step game is determined by the recursive relation

$$E(L_n) = 1 + \left(\frac{2}{4-M_n}\right)^2 \cdot E(L_{n-1}), \quad n=2,\ldots, \quad E(L_1) = 1\,.$$

Similarly the second moment is given by

$$E(L_n^2) = 1 + \left(\frac{2}{4 - M_n}\right)^2 \cdot (2 \cdot E(L_{n-1}) + E(L_{n-1}^2)), \quad n = 2, \ldots, \quad E(L_1^2) = 1.$$

Therewith we obtain for the $E(L_n)$ and the standard deviation $Std(L_n)$ for all n

$$E(L_n) < 1.8 \qquad \text{and} \qquad Std(L_n) < 1.2;$$

we see that for $n \geq 4$ more than two rounds of the game will hardly be played. In fact, for $n \geq 4$ the probability that the absurd strategy combination has to be accepted as the solution of the game is less than 0.03 (which is much smaller than the probability 0.16 in the 1-step game).

4 Generalizations

It looks strange that we talked about an absurd strategy combination even though another combination – both go separately to their preferred events – has the same payoffs. The reason for the set of payoffs given in Figure 1 was that we wanted to maintain the original payoffs by Luce and Raiffa and by Rapoport. One can see immediately, however, that our results, in particular the recursive formulae, would be the same if we chose worse payoffs to both players in the absurd strategy combination than in the one in which both go separately to their preferred events.

The situation changes completely, if we assume altruistic behavior of both players. Let us consider Figure 2.

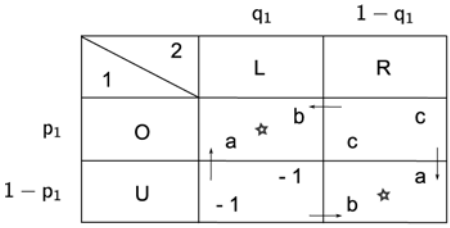


Fig. 3 Normal form of the Battle of Sexes paradigm with generalized strategies and payoffs.

In order to maintain the equilibrium structure of Figure 1 we assume $-1 < a$ and $c < b$. Then the mixed equilibrium payoffs are

$$\frac{ab + c}{a + b - c + 1}. \tag{6}$$

If we proceed as before, in case the absurd combination is realized, we enter the 2nd step of the game with payoffs in the lower left box of the reduced normal form given by (6).

But now the only equilibrium of the game in case of

$$a < \frac{a\,b+c}{a+b-c+1} \qquad \text{or equivalently} \qquad a < c \tag{7}$$

is precisely the absurd combination of strategies! In this case we have because of $c < b$ and (7) $a < b$, that is, in the battle of sexes game the man indicates less pleasure in boxing than his wife and the wife less pleasure in the ballet than her husband if they visit these events together. Thus, it is not surprising that these payoffs may lead to the strange result indicated above.

This way one may interpret the meetings of the French and German delegations mentioned in the second section: Since both delegations behaved altruistically, and since they played in the first round the absurd strategy combination, they played in the second round the same namely the only equilibrium strategy combination!

5 Concluding Remarks

Only in rare cases it is possible to describe a serious conflict with the help of a normal form game, not only because then unrealistic strategy combinations may have to be considered. If this happens however, then these games may at best describe the initial conflict situation for which no one-shot solution can be given, if no mediator is accepted and no bargaining solution whatsoever can be agreed upon.

Thus one has to model this situation as a game over time. It depends on the concrete conflict if it is appropriate to describe the effort for its solution as a repeated game (in the sense we did it), or if discounting elements have to be introduced, or new moves eventually of a random nature. In any case, as the Sudan conflict showed it may become very difficult to obtain all information necessary for such a more realistic modelling.

References

1. R. Avenhaus. French-German Official Meeting. *PINPoints, IIASA Laxenburg*, 26(1): 11, 2006.
2. R. Avenhaus and Th. Krieger. A Fresh Look on the Battle of Sexes Paradigm — Proofs and Extensions. *Technischer Bericht, Fakultät für Informatik*, 2010.
3. R. Luce and H. Raiffa. *Games and Decisions*. John Wiley & Sons, New York, 1957.
4. G. Owen. *Game Theory*. Academic Press, New York, 2nd edition, 1982.
5. A. Rapoport. *Fights, Games and Debates*. The University of Michigan Press, Ann Arbor, 1974.
6. J. Smilovitz. Identity and Mediation. In M. Anstey, P. Meertz, and I. W. Zartman, editors, *Reducing Identity Conflicts and Preventing Genocide*. 2011.
7. E. van Damme. *Stability and Perfection of Nash Equilibria*. Springer-Verlag, Berlin Heidelberg, 1987.

A Generalization of Diamond's Inspection Model: Errors of the First and Second Kind

Thomas Krieger

Abstract In many material processing and storing plants an inspector performs during some reference time interval a number of inspections because it can not be excluded that the operator acts illegally by violating agreed rules, e.g., diverts precious or dangerous material. In many situations it is necessary to allow incomplete detections and the possibility of false alarms.
In this contribution a model for unobservable interim inspections due to Diamond is extended to include first and second kind errors in case of one or two unannounced interim inspections per reference time interval. In both cases a Nash equilibrium is determined and its properties are discussed. This contribution is also intended to bring Diamonds brilliant idea of solving those kind of games back to awareness.

1 Introduction

A number of years ago Diamond, see [4], derived a general solution for the problem of determining a randomized inspection strategy which optimizes the time between the beginning of an event or activity which is not in line with the specification of some process, or even illegal according to some agreement. In his model, Diamond assumed certain detection and no possibility of false alarms for the interim inspections, and also that the operator will behave illegally with certainty. These assumptions were consistent with modelling the problem as a zero-sum game with time from the start of the illegal activity to its detection as payoff to the operator.

In the context of a nuclear safeguards inspection regime, see e.g., [5], or more general of inspecting precious or dangerous material, it is necessary, e.g., for cost reasons, to allow incomplete detection and the possibility of false alarms. This happens if *Variable Sampling Procedures* are applied, see e.g., [1].

Thomas Krieger
ITIS GmbH an der Universität der Bundeswehr München, Fakultät für Informatik, Werner-Heisenberg Weg 39, 85579 Neubiberg, e-mail: thomas.krieger@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_13,

Since false alarms are disadvantageous for both parties, the zero-sum assumption is then no longer tenable. Moreover, the alternative of legal behavior on the part of the operator likewise cannot be treated within a zero-sum game model.

Therefore, a general non-cooperative game between inspector and operator is formulated in which the players' utilities reflect their priorities over a reference time interval. The inspector prefers deterrence (that is, legal behavior on the part of the operator) to detection of an illegal activity; the operator prefers not detected illegal behavior to detected illegal behavior, and both protagonists wish to avoid false alarms. Both players decide on their strategies in advance, i.e., at the beginning of the reference time interval.

2 The Model

To be specific we consider a single inspected object, for example a nuclear facility subject to verification in the framework of an international treaty, and the reference time $[0,T]$. The illegal activity may be the diversion of nuclear material for weapons manufacturing purposes. We assume that a thorough and ambiguous inspection takes place at the end of the reference time interval which will detect an illegal activity with certainty if one has occurred. In addition less intensive and strategically placed *interim* inspections take place which are intended to reduce the time to detection below the length of the reference time interval T. The interim inspection will detect a preceding illegal activity, but with some probability smaller than one. In keeping the common notation, we call this probability $1-\beta$, where β is the probability of an error of the second kind, or *non-detection probability*. If the interim inspection is not preceded by an illegal activity, there is a corresponding probability of an error of the first kind, the *false alarm probability* α.

The preferences of the players (operator,inspector) are taken to be as follows:

$(0,0)$ for legal behavior over reference time interval, and no false alarm,
$(-f,-g)$ for legal behavior over reference time interval, and a false alarm
$(d\,\Delta t - b, -a\,\Delta t)$ for the detection of the illegal activity after elapsed time $\Delta t \geq 0$,

where $0 < g < a$ and $0 < f < b < d\,T$, see [7]. Thus the preferences are normalized to zero for legal behavior without false alarms, and the loss (profit) to the inspector (operator) grows proportionally with the time elapsed to detection of the illegal activity. A false alarm is resolved unambiguously with time independent costs $-g$ to the inspector and $-f$ to the operator, whereupon the game continuous. Note that, if $d\,T - b < 0$, the operator will behave legally even if there are no interim inspections at all. Since the interim inspections introduce false alarm costs for both parties, there would be no point in performing them.

2.1 The Game with One Unannounced Interim Inspection

Let, as mentioned, the reference time interval be $[0,T]$. The operator starts his illegal activity at time $s \in [0,T)$ and the inspector chooses his interim inspection time point $t_1 \in (0,T)$. The operator's payoff is then given by the payoff kernel

$$Op_1(s,t_1) := \begin{cases} d\,[(t_1-s)\,(1-\beta)+(T-s)\,\beta]-b & : \quad 0 \le s < t_1 < T \\ d\,(T-s)-f\,\alpha-b & : \quad 0 < t_1 \le s < T \end{cases} . \quad (1)$$

The payoff to the inspector $In_1(s,t_1)$ is obtained from (1) by replacing d by $-a$, f by g and setting $b=0$. The justification of this formula and further model assumptions are given in [2] and [7]. The most important assumption is, that any illegal activity is guaranteed to be detected at the end of the reference time interval.

This game does not have a Nash equilibrium in pure strategies. Therefore, we have to consider mixed strategies, i.e., distribution functions on $\mathbb{R}$, see, e.g., [6]. Let $Q(s)$ be the probability of diversion occurring at time s or earlier and let $P(t)$ be the probability of an inspection having taken place at time t or earlier. Using Lebesque-Stieltjes integrals, see, e.g., [3], we define the expected payoff to the operator by $Op_1(Q,P) := \int_{[0,T)} \int_{(0,T)} Op_1(s,t)\,dQ(s)\,dP(t)$ and to the inspector by $In_1(Q,P) := \int_{[0,T)} \int_{(0,T)} In_1(s,t)\,dQ(s)\,dP(t)$. A mixed strategy combination (Q^*,P^*) constitutes a Nash equilibrium if and only if $Op_1(Q^*,P^*) \ge Op_1(Q,P^*)$ for all Q and $In_1(Q^*,P^*) \ge In_1(Q^*,P)$ for all P.

Infinite games with discontinuous payoff kernels, such as the games in this paper, may have no Nash equilibrium at all, see, e.g., [5]. Fortunately, it can be shown, that for the games discussed here, at least one Nash equilibrium exists. This is formulated for the case of one unannounced interim inspection in

Theorem 1 *Let $h_1(x)$ be the solution of the differential equation*

$$h_1'(x) = (1-\beta)\,h_1(x) - \alpha\,f/d \qquad \textit{with} \qquad h_1(0) = A_1\,, \quad (2)$$

where $A_1 > 0$ is determined by the condition $h_1(1) = T$. Furthermore, let the test procedure be unbiased, i.e., $\alpha+\beta < 1$. Then the normal form game for one unannounced interim inspection has the following equilibrium: The operator chooses his time s for starting the illegal activity from $[0,T)$ according to the distribution function

$$Q_1^*(s) = \begin{cases} 0 & : \quad s \in (-\infty,0) \\ \dfrac{c_1}{T}\left[h_1\left(\displaystyle\int_0^s \frac{d\xi}{h_1'(h_1^{-1}(T-\xi))+\gamma}\right) - \frac{\alpha}{1-\beta}\,\frac{f}{d}\right] & : \quad s \in [\underbrace{T-h_1(1)}_{=0}, T-h_1(0))\,, \\ 1 & : \quad s \in [T-h_1(0),\infty) \end{cases} \quad (3)$$

with $\gamma := \alpha\,(g/a+f/d)$ and

$$c_1 = \left(\frac{1}{T} \left[h_1 \left(\int_0^{T-h_1(0)} \frac{d\xi}{h_1'(h_1^{-1}(T-\xi)) + \gamma} \right) - \frac{\alpha}{1-\beta} \frac{f}{d} \right] \right)^{-1}.$$

Furthermore, let u be a realization of a uniformly distributed random variable U on $[0,1]$. *Then the inspector chooses his equilibrium inspection time point* t_1^* *according to*

$$t_1^* = T - h_1(1-u). \tag{4}$$

The corresponding equilibrium payoffs are given by

$$Op_1^* = d\,h_1(0) - f\,\alpha - b \quad \textit{and} \tag{5}$$

$$In_1^* = (-a) \left[T - (1-\beta)\,h_1(0) - \int_{[0,T)} s\,dQ_1^*(s) \right]. \tag{6}$$

Proof. The proof of this Theorem can be found in [7]. □

Let us discuss this result: First, Diamond's brilliant idea was to describe the equilibrium inspection time point as a function of a uniformly distributed random variable U on $[0,1]$.

Second, if one solves (2), the equilibrium strategies as well as the equilibrium payoffs can be given explicitly, see [2]. Here, the implicit representation (3) - (6) using $h_1(x)$ is preferred, since one can more easily guess how this result can be generalized to the game with two unannounced interim inspections. For later purposes we give the explicit expression of the operator's equilibrium payoff (5):

$$Op_1^* = d \left(e^{-(1-\beta)} \left(T - \frac{\alpha}{1-\beta} \frac{f}{d} \right) + \frac{\alpha}{1-\beta} \frac{f}{d} \right) - f\,\alpha - b. \tag{7}$$

Third, from (4) one sees that after time point $T - h_1(0)$ neither the illegal activity is started nor the inspection is performed. The reason for that surprising effect is, that detection is guaranteed to occur at the end of the reference time interval and the operator will not start his illegal activity too late. Furthermore it is interesting to notice that the operator will start his illegal activity with positive probability at time point 0, since with (3) it is $Q_1^*(0) > 0$.

Fourth, in case of $\alpha = 0$, i.e., only *Attribute Sampling Procedures* are applied, see e.g., [1], the mid line of formula (3) can be simplified to $[h_1(1 - h_1^{-1}(T-s))]/T$, which shows the immense consequences in the complexity of the Nash equilibrium if one introduces false alarms. Furthermore, the equilibrium strategies of both players do *not* depend on the payoff parameters, which is very interesting for practical implications. The equilibrium payoff to the operator is with (7) given by $Op_1^* = d\,h_1(0) - b = d\,T\,e^{-(1-\beta)} - b$. It can be shown, see [7], that in case of $\alpha = 0$ the inspector's equilibrium payoff is given by $In_1^* = -a\,h_1(0)$, i.e., has the same structure as (5).

Fifth, in case of $\alpha = \beta = 0$, $d = a = 1$ and $b = 0$ our game degenerates to Diamond's zero-sum game, where the payoff to the operator can be directly interpreted as *detection time*, i.e., the time between start and detection of the illegal activity. With (7) the optimal expected detection time is $Op_1^* = h_1(0) = T\,e^{-1} \approx 0.3679\,T$.

2.2 The Game with Two Unannounced Interim Inspections

Let us now analyze the game with two unannounced interim inspections. Again let $s \in [0,T)$ be the time point for starting the illegal activity and let t_1 and t_2 with $0 < t_1 < t_2 < T$ be the time points for the interim inspections. Then the payoff kernel for the operator is given by

$$Op_2(s,(t_1,t_2)) = \begin{cases} d\Big[(1-\beta)(t_1-s)+ \\ \qquad +\beta(1-\beta)(t_2-s)+\beta^2(T-s)\Big]-b & 0 \le s < t_1 < t_2 < T \\ d\Big[(1-\beta)(t_2-s)+\beta(T-s)\Big]-\alpha f-b & : \quad 0 < t_1 \le s < t_2 < T \\ d(T-s)-2\alpha f-b & : \quad 0 < t_1 < t_2 \le s < T \end{cases} \tag{8}$$

Again, the payoff to the inspector $In_2(s,(t_1,t_2))$ is obtained from (8) by replacing d by $-a$, f by g and setting $b = 0$. The justification of this formula is given in [7].

Theorem 1 shows that the equilibrium strategy $Q_1^*(s)$ of the operator is rather complicated. Since, of course, it is even more complicated for two inspections, it is omitted in the following Theorem. This appears to be justified even more so, since for applications especially the equilibrium strategy of the inspector is asked for. The equilibrium payoff of the operator, however, is interesting again, since with its help conditions for deterring the operator from illegal behavior are given.

Theorem 2 *Let $h_1(x)$ and $h_2(x)$ be the solution of the system of differential equations*

$$\begin{aligned} h_1'(x) &= (1-\beta)h_1(x) - \alpha f/d \qquad \textit{with} \qquad h_1(0) = A_2 \\ h_2'(x) &= (1-\beta)h_2(x) - (1-\beta)^2 h_1(x) - \alpha f/d \qquad \textit{with} \qquad h_2(0) = h_1(1), \end{aligned}$$

where $A_2 > 0$ is determined by the condition $h_2(1) = T$. Furthermore, let the test procedure be unbiased, i.e., $\alpha + \beta < 1$. Let the operator's payoff parameters f and d be such that $h_1(x)$ and $h_2(x)$ are monotonely increasing functions in x. Then in the normal form game for two unannounced interim inspections there exists a Nash equilibrium with the following property: The inspector chooses his equilibrium inspection time points (t_1^*, t_2^*) according to

$$t_1^* = T - h_2(1-u) \qquad \textit{and} \qquad t_2^* = T - h_1(1-u), \tag{9}$$

where u is a realization of a uniformly distributed random variable U on $[0,1]$. The equilibrium payoff to the operator is given by

$$Op_2^* = d\,h_1(0) - 2f\alpha - b.$$

Proof. The proof of this Theorem can be found in [7]. □

Again, Diamond's idea enables us to describe the equilibrium inspection time points as a function of a uniformly distributed random variable U on $[0,1]$.

As in the previous section, one sees with (9) that the operator resp. the inspector starts his illegal activity resp. performs his inspections not later than $T - h_1(0)$. The

explanation of this effect is the same as given there. Again, the operator will start his illegal activity with positive probability at time point 0, see [7].

The requirement that $h_1(x)$ and $h_2(x)$ have to be monotone increasing functions, is always fulfilled for $\alpha = 0$, see [7]. For $\alpha > 0$ this requirement does not pose serious limitations to the operator's payoff parameters, see [7].

In case of $\alpha = \beta = 0$, $d = a = 1$ and $b = 0$, i.e., the case of Diamond's zero-sum game, the optimal expected detection time is given by $Op_2^* = h_1(0) = T/(e(e-1)) \approx 0.2141\,T$, i.e., smaller than in the case of one unannounced interim inspection, as expected.

3 Conclusions

In line with Diamond's model it is assumed that the operator will behave illegally with certainty. In general, the main goal of an inspection agency is, however, to deter the operator from behaving illegally. If one wants to model also the possibility of legal behavior, a larger game has to be considered, in which the legal behavior becomes a pure strategy of the operator. It can be shown that the operator is deterred from behaving illegally if and only if $b/d > h_1(0)$ holds. Note that $h_1(0)$ is different for the cases of one resp. two unannounced interim inspections. This way we get lower bounds on the ratio of sanctions to incentives b/d for deterrence, see [7].

Acknowledgements The author would like to thank Rudolf Avenhaus, Universität der Bundeswehr München, for helpful suggestions and active support. The ITIS GmbH has providing the possibility to work on the problems discussed here.

References

1. R. Avenhaus and M. J. Canty. *Compliance Quantified — An Introduction to Data Verification*. Cambridge University Press, Cambridge, UK, 1996.
2. R. Avenhaus, M. J. Canty, and K. Sohrweide. Extension of a Model for Timely Inspection. In *Proceedings of the 25th Annual Meeting,ESARDA Syplosium in Safeguards and Nuclear Material Management, Stockholm, 13–15 May, 2003*, JRC Ispra, Italy 2003.
3. M. Carter and B. van Brunt. *The Lebesque-Stieltjes Integral*. Springer-Verlag, New York, 2000.
4. H. Diamond. Minimax Policies for Unobservable Inspections. *Mathematics of Operations Research*, 7(1): 139–153, 1982.
5. IAEA. Model Protocol Additional to the Agreement(s) Between State(s) and the International Atomic Energy Agency for the Application of Safeguards. INFCIRC/540. Technical report, IAEA, Vienna, 1997.
6. S. Karlin. *Mathematical Methods and Theory of Games, Programming, and Economics — Volume II*. Addison-Wesley, Reading, Massachusetts, 1959.
7. Th. Krieger. *Inspektionen mit Fehlern erster und zweiter Art (Inspections with errors of the first and second kind (in German))*. forthcoming, 2011.
8. G. Owen. *Game Theory*. Academic Press, New York, 2nd edition, 1982.

I.3 Managerial Accounting

Chair: Prof. Dr. Matthias Amen (Universität Bielefeld)

Accounting figures affect management decisions. Traditionally we separate between financial and management accounting. Usually management accounting is more connected to the objective-oriented allocation of inputs with the focus on either internal decision making or managerial behavior. Financial accounting provides information for external investors.

Today, both accounting perspectives are strongly linked. According to the management approach in IFRS, financial statements have to report some figures originally used only for internal purposes, e.g. segment or risk reporting. On the other hand, investors' information requirements force the management to consider financial reporting figures in the internal control system.

Therefore, the topic considers not only decision making or behavioral aspects in all fields of classical management accounting, but also the link to financial reporting as well as internal and external auditing. Especially classical operations research approaches in Quantitative Accounting are addressed, including optimization, simulation, forecast and analysis methods.

Case-Mix-Optimization in a German Hospital Network

Katherina Brink

Abstract The dire financial situation of the German health care system, where the insurance companies and the federal states act as the paying authorities for the operative costs and investment costs, enhances the need for hospitals to increase their efficiency and cut costs. This has resulted in a rising number of co-operations in the hospital market. To exploit the synergetic potential deriving from a co-operation with other hospitals in a region, the amount and type of cases treated at the hospitals play an important role. The decision regarding the composition and allocation of the case-mix among the members of a hospital network has to be made under consideration of logical, legal, organizational, and capacity restrictions.

In this paper, a management accounting model is introduced and described on the basis of already existing research as an aid for this decision making process. Additionally, modifications necessitated by changes in influencing factors are presented.

1 Current Situation and Research Objective

Due to financial problems of the statutory health insurance companies, which account for approx. 82% of the operating costs, and the desolate public budget of the federal states, which are responsible for financing investments (e.g., in the infrastructure), hospitals are under high pressure to cut costs and improve their efficiency. As early as 1972, Dietrich Adam addressed the problem of rising costs and weak profitability in the hospital market and the necessity of economically sound behaviour [1]. Five years later, in 1977, the German Bundestag passed the first law to diminish costs. The main focus of reforms adopted since then focused on hospital financing and especially on limiting and reducing the expenses of the insurance companies. One of the last major reforms - the introduction of the German DRG

Katherina Brink
Universität Bielefeld, Universitätsstr. 25, D-33615 Bielefeld, e-mail: katherina.brink@uni-bielefeld.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_14,

System (Diagnosis Related Group) in 2003 as a fixed pricing system - increased the pressure to economize. It enhanced the reduction of a patient's length of stay and led to a rise in co-operations within the market. The number of horizontal co-operations between hospitals increased, for example, by 53% between 2003 and 2007. In order to exploit the synergetic potentials for cost reduction and revenue increase, resulting from a co-operation, decisions have to be made concerning the intensity rate of the co-operation (ranging from a lose co-operation form to a merger), the strategy to be followed (e.g., specialization), and which case-mix the hospitals within the network will offer.

The research objective of this paper results from these considerations. While constructing a hospital network out of existing hospitals, the question is how the different cases can be allocated across the different departments and locations in the network in compliance with capacity, as well as legally and organizationally requested restrictions. As efficiency and cost reduction are the main goals, the problem is solved with recourse to management accounting tools.

2 Problem Setting

Each hospital location (s) consists of a number of departments (n) where a number of DRG-cases (i) is being treated. The case-mix of the hospitals within a region overlaps due to the historically grown and legally given restrictions as well as the preferences of physicians to assign patients to a certain hospital. As an illustrative example three hospitals are introduced [cf. Figure 1].

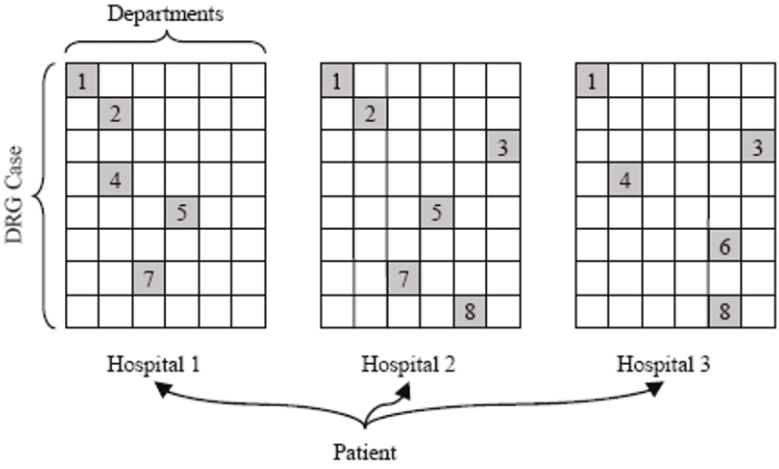


Fig. 1 Non-cooperating Hospitals

Each of the three hospitals consists of six departments which treat eight DRG cases each. DRG 1 is treated by all three hospitals, whereas DRG 6 is only treated by hospital 3. Overlaps in the case-mix between the hospitals exist in case of DRG 2, 3, 4, 7, and 8. The hospitals compete with other hospitals for the patients in the region to cover their costs. Due to the public mandate to ensure an overall treatment of the patients, hospitals also have to accept less profitable cases. Regulations make it difficult for the hospitals to change their case-mix short-term. While the number of beds per department is defined by the federal states in their hospital plan, the amount and type of the DRGs to treat (case-mix) are set during the course of budget negotiations with the insurance companies. As the case revenue as well as the average costs is given, the case-mix can be used to determine the profitability of a hospital.

In case of an establishment of a horizontal network, the hospitals are able to adjust their case-mix more easily, provided that the total amount of defined cases does not change [cf. Figure 2].

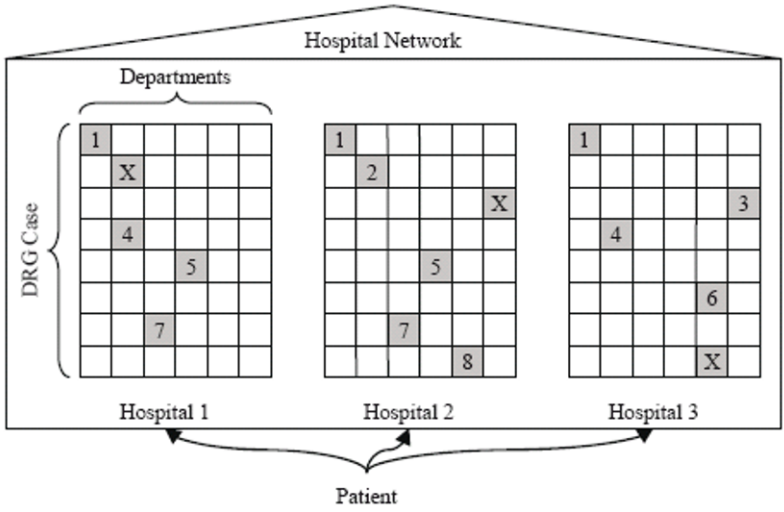


Fig. 2 Horizontal Hospital Networks

In this example the DRGs 2, 3, and 8 are only treated at one location after the case-mix of the hospital network has been optimized. Due to a high number of DRG 1 cases in the region it is still treated at all three. The restructuring leads to diminished costs and higher profits and thereby insures a sustainable existence in the hospital market and an increase of profitability. To decide which DRG case to treat at the hospitals within the network, or only at one or two, the decision maker can revert to a quantitative planning model with binary variables. The inability to store services and the uncertainty of demand make planning and allocation thereby even more difficult. In order to simplify the model, demand and capacity terms are set

as deterministic parameters. The allocation problem can be classified as an NP-hard optimization problem [5].

3 Basic Short-Term Approach

Research has developed optimization models with an economic perspective to facilitate the regional hospital planning by federal states or the management of individual hospitals. From a management accounting perspective the research contributions made by Harfner (1999) and FleSSa (2006) are of particular interest (cf. [6] and [4]). With some adjustments, e.g. of restrictions, they form the basis for the described approach. Both devised a short-term deterministic static model which is based on the multi-level contribution income statement with the aim to find a case-mix portfolio that maximizes profit. The described model relies on the following assumptions:

- Profit orientation as primary goal
- Rational decision makers
- Focus on case-mix
- Focus on operating costs and neglecting of investment costs
- Costs and revenue per case, as well as capacity limits are given

The target function can therefore be formulated as such[1]

$$P = \sum_{i=1}^{I}\sum_{s=1}^{S}\left(p_i - k_{is}^{dir}\right) \times x_{is} - \sum_{i=1}^{I}\sum_{s=1}^{S} K_{is}^{fix} \times \delta_{is} - \sum_{n=1}^{N}\sum_{s=1}^{S} K_{ns}^{fix} \times \beta_{ns} - \sum_{s=1}^{S} K_{s}^{fix} \times \tau_s \rightarrow max$$

In accordance to the statements above the following restrictions have to be regarded in the allocation model

- *Logical Restrictions*: These restrictions enable the allocation variables δ_{is}, β_{ns}, and τ_s to become "1" or "0". They define if a case is treated at a department at a certain location within the hospital network. If a case is allocated to a location, i.e. the variables become "1", the relevant fix costs are allocated to the case K_{fix}^{is}, department K_{fix}^{ns}, and location K_{fix}^{s}.
- *Legal Restrictions*: The German hospital market can be characterized by a high rate of regulations concerning bed numbers per department, case mix and budget. As the majority of hospitals in the German market are included in the hospital plan of the federal states and therefore are subject to these regulations they are

[1] p_i: revenue per DRG-case i; k_{is}^{dir}: direct costs of a DRG-case, e.g. material costs for medicine or an operation; x_{is}: amount of case i at location s; K_{is}^{fix}: DRG fix costs, e.g. personnel costs for a midwife; δ_{is}: binary variable that becomes "1" if case i is allocated to location s; K_{ns}^{fix}: department fix costs, e.g. personnel costs for the staff allocated to the specific department; β_{ns}: binary variable that becomes "1" if department n is allocated to location s; K_{s}^{fix}: location fix costs; e.g. personnel costs for the rest of the hospital staff, material costs for other hospital functions; τ_s: binary variable that becomes "1" if location s exists.

regarded here as well. The relevant legal texts in this matter are the Krankenhausentgeltgesetz (Hospital Renumeration Law) and the Sozialgesetzbuch V (Social Law Code No. 5).

- *Organizational Restrictions*: The more hospitals are specialized the higher is the risk that the personnel-in-training cannot go through all necessary departments and procedures on one location but have to go to another. Additionally, more patients are going to be neglected if hospitals are too specialized [2]. The organizational restrictions take these circumstances into account.
- *Capacity Restrictions*: The capacity restrictions concerning personnel, operating rooms, and beds are the largest set of restrictions. The bed capacity is linked to the legal restrictions and the length of stay of each patient. Each patient requires a certain amount of personell time for diagnosis, therapy, operative procedures and care. The individual time requirements depend on the characteristics of each case.

The restrictions not only formally constrain the solution space but help to integrate market and problem specific characteristics. The assumption of deterministic values leads to a small set of alternive courses of action. However, the adjustments in the case-mix per location can only be changed long-term. Thereby, temporal aspects like the adjustment processes of parameters, possible synergetic effects caused by the co-operation, as well as changes in demand and prices, gain importance. Modifications become necessary which are based on the introduced model.

4 Modifications

The majority of prior research regarding the problem of case-mix optimization (management perspective) formulate a short-term approach in a one hospital setting. FleSSa (2006) and Harfner (1999) are examples of a model solution for a horizontal hospital network. However, as stated above, they do not regard the impact of time on prices, costs, capacity supply and demand. In a long-term perspective former fix parameters become variable resulting in a higher scope of solutions and possibilities. So neglecting temporal issues has a deep impact on the final decisions.

When it comes to the formulation of a long-term model two possible modifications come into mind: a static or a dynamic model [7]. Only a few long-term/ multi-periodic static models exist so far in (regional) hospital planning research (e.g. [3] and [9]). The authors integrate the factor time as an additional index. However, each period is optimized separately, i.e. without regarding the interdependencies between the periods. The timing is defined as discrete.

As opposed to a static model, a dynamic one regards the adjustment process of the parameters over time. Possible dynamic effects to be considered are the learning curve as well as the change in the demand structure and price developments. The learning effect, e.g., leads to a decline in capacity demand for operative procedures. Capacity becomes available for more operations. A reduction in length of stay also reduces the capacity demand per case. The hospital is able to accept more patients.

The marginal costs per case decline. Dynamic models only exist for services in general so far (e.g. [8]), but not in a hospital context. The basic approach as well as the modifications assume deterministic values. However, they can also be set as stochastic for further modifications.

5 Conclusion

As one can see from the discussions above further research is required. More competition and efficiency oriented regulations in the hospital market simplify the application of management concepts - especially in favor of management accounting tools as they focus on efficiency and profitability. In the existing research the long-term perspective has been neglected so far. Thereby, in this research project the basic short-term approach will be extended by long-term static and/or dynamic properties. The aim is to integrate time as an influencing factor and possibly consider uncertainty via stochastic values. Additionally, a sensitivity analysis can help to identify the impact certain parameters and/or restrictions have on hospital profits. In the end, this management accounting tool can be used as an aid for budget negotiations as well as for integration management.

References

1. A. Dietrich. *Krankenhausmanagement im Konfliktfeld zwischen medizinischen und wirtschaftlichen Zielen. Eine Studie über Möglichkeiten zur Verbesserung der Strukturorganisation und des Entscheidungsprozeses in Krankenhäusern.* Gabler, Wiesbaden, 1972.
2. I. Duncan and B. Nobel. The Allocation of Specialties to Hospitals in a Health District. *The Journal of the Operational Research Society*, 30: 953–964, 1979.
3. S. Fleßa. Ressourcenallokation und Zeitpräferenz in der Gesundheitsdistriktplanung von Entwicklungsländern. *OR Spektrum*, (23): 203–222, 2001.
4. S. Fleßa, B. Ehmke, and R. Herrmann. Optimierung des Leistungsprogramms eines Akutkrankenhauses - Neue Herausforderungen durch ein fallpauschaliertes Vergütungssystem. *BFuP*, 58(6): 585–599, 2006.
5. M. Garey and D. Johnson. *Computers and intractability. A guide to the theory of NP-completeness*. Freeman, New York, 24th edition, 2003.
6. A. Harfner. *Spezialisierungs- und Konzentrationsprozesse im deutschen Krankenhauswesen bei einem fallbezogenen Finanzierungssystem. Eine quantitative Analyse mit Hilfe computergestützter Szenarienrechnung*. Forschungsgruppe Medizinökonomie Universität Erlangen-Nürnberg (Arbeitsbereicht Nr. 99,2), Nürnberg, 1999.
7. A. Scholl. *Robuste Planung und Optimierung. Grundlagen - Konzepte und Methoden - experimentelle Untersuchung*. Physica Verlag, Heidelberg, 2001.
8. M. Schweitzer. *Taktische Planung von Dienstleistungskapazitäten. Ein integrativer Ansatz.* Duncker & Humbolt, Berlin, 2003.
9. C. Stummer, K. Doerner, A. Focke, and K. Heidenberger. Determing Location and Size of Medical Departments in a Hospital Network: A Multiobjective Decision Support Approach. *health Care Management Science*, 7: 63–71, 2004.

I.4 Financial Modelling and Numerical Methods

Chair: Prof. Dr. Daniel Rösch (Leibniz Universität Hannover)

Financial decision making problems and approaches for their solutions by financial modeling and numerical methods have become likewise more and more important and rigorous during the last decade. This includes numerical solutions to complex pricing models for financial instruments, optimization and simulation in hedging and risk management, and quantitative and numerical analysis relevant for finance.

We welcome submissions of theoretical, empirical and practical papers on all aspects of financial modeling including models, computational and numerical methods and applications.

New Insights on Asset Pricing and Illiquidity

Axel Buchner

Abstract Many important asset classes are illiquid in the sense that they cannot be traded. When including such illiquid investments into a portfolio, portfolio decisions take on an important dimension of permanence or irreversibility. Using a continuous-time model, this extended abstract shows that this irreversibility leads to portfolio proportions being stochastic variables over time as they can no-longer be controlled by the investor. Stochastic portfolio proportions have major implications since they can change portfolio dynamics in a fundamental way. In particular, it is shown that stochastic proportions implied by illiquidity increase overall portfolio risk. Interestingly, this effect gets more pronounced when the return correlation between the illiquid and liquid asset is low, i.e., the increase in portfolio risk caused by illiquidity is inversely related the return correlation of the illiquid and liquid assets.

1 Introduction

Many important classes of assets are illiquid in the sense that they cannot be traded. The impact of illiquidity on expected returns and asset prices has been the subject of numerous theoretical and empirical studies.[1]

The model developed in this extended abstract complements to the existing literature on illiquidity and asset pricing. The main idea is that including an illiquid asset into a portfolio of liquid assets has major implications since it can change the dynamics of the portfolio in a fundamental way. When including assets that cannot be sold into a portfolio, portfolio decisions take on an important dimension of permanence or irreversibility. This irreversibility is reflected by the fact that an investor can no longer rebalance the portfolio freely after market movements. In this

Dr. rer. pol. Axel Buchner
Department of Financial Management and Capital Markets, Technical University of Munich, Arcisstrasse 21, 80333 Munich, e-mail: axel.buchner@wi.tum.de

[1] Important theoretical and empirical work in this area includes [1, 2, 4, 5].

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_15,

sense, the definition of illiquidity adopted here is that it is the foregone opportunity to control portfolio proportions over time when illiquid and liquid assets are combined in a portfolio. This foregone opportunity raises a number of key asset pricing issues. In particular, how does it affect expected returns and risk of a portfolio if asset proportions cannot be fully controlled?

To address these issues, this extended abstract examines the asset pricing implications of illiquidity within a continuous-time model. In this framework, the liquid asset can always be traded. The illiquid asset can always be bought on the market, but cannot be sold once it has been acquired. However, the illiquid asset provides a liquid "dividend". This liquid dividend is assumed to be reinvested into the liquid asset in the model. In addition, it is assumed that parts of the liquid asset are liquidated at each instant in time and are invested into the illiquid asset. This modeling approach results is a circular flow of capital between the investor's sub-portfolio invested liquid and illiquid that captures well the typical process how investments into illiquid assets are made over time.

This extended abstract makes two novel contributions to the existing literature on asset pricing and illiquidity. The first major implication is to show that illiquidity results in portfolio weights being stochastic processes over time. This is an intuitive result, as portfolio weights are no longer under the full control of the investor when parts of the assets in the overall portfolio cannot be sold. The economic relevance of this result stems from the fact that stochastic portfolio proportions have important implications for the return dynamics of a portfolio.

In this sense, the second contribution of the extended abstract is to analyze how stochastic portfolio proportions implied by illiquidity affect the expected return and risk of a portfolio. Specifically, it is shown that the risk of a portfolio increases when parts of the assets in the portfolio cannot be sold. In order to illustrate this effect, it is demonstrated that the efficient frontier derived from the model always lies within the space spanned by the efficient frontier that arises from a traditional mean-variance analysis in the reference case when all assets are liquid. That is, the same expected portfolio return can only be achieved with higher levels of risk when parts of the portfolio are illiquid. Thereby, the increase in risk is inversely related to the return correlation, i.e., a lower correlation results in a higher additional risk of the portfolio. It is a main result from the traditional mean-variance theory that portfolio risk decreases with decreasing levels of the return correlations of the assets. This result still continues to hold under the developed model setting. However, it is extended here as it is shown here that illiquidity particularly increases portfolio risk when the return correlation between the liquid and illiquid assets is low.

2 The Model

We start by considering an investor's portfolio that consists of two components: sub-portfolio L and I. Sub-portfolio L denotes the capital that is invested in liquid assets, sub-portfolio I the capital invested in illiquid assets. From this, it follows that the

value P_t of the investor's total portfolio at any time $t \geq 0$ can be expressed in terms of the two sub-portfolio values, i.e., $P_t = L_t + I_t$ holds.

The flow of capital between the two sub-portfolios is governed by two model parameters. These are the dividend yield of the illiquid assets $\delta > 0$, and the investment rate $\nu > 0$. Thereby, it is assumed that capital contained in L moves to I according to ν, the rate of new investments in illiquid assets relative to the sub-portfolio value of the liquid assets L. Simultaneously, capital moves from I to L according to δ, the rate of dividends that are reinvested in liquid assets relative to the sub-portfolio value of illiquid assets I. Both structural model parameters are, for simplicity, assumed to be positive constants over time. Uncertainty is introduced into the model by assuming that liquid and illiquid asset returns R_L and R_I are normally distributed with constant means and standard deviations, i.e., $R_L \sim N(\mu_L, \sigma_L^2)$ and $R_I \sim N(\mu_I, \sigma_I^2)$. Then, the circular flow of capital between the sub-portfolios can be described by the system of stochastic differential equations

$$dL_t = L_t(\mu_L - \nu)dt + L_t\sigma_L dB_{L,t} + I_t\delta dt, \tag{1}$$
$$dI_t = I_t(\mu_I - \delta)dt + I_t\sigma_I dB_{I,t} + L_t\nu dt, \tag{2}$$

where $B_{L,t}$ and $B_{I,t}$ are standard Brownian motions with correlation $d\langle B_{L,t}, B_{I,t}\rangle = \rho dt$. The continuous-time stochastic differential equations (1-2) represent a *two-dimensional* stochastic system, driven by a *two-dimensional* Brownian motion. The *drift* and *diffusion* coefficients of the stochastic system (1-2) are *homogeneous* functions of *degree one* in the state variables L and I. Fortunately, this allows to analytically describe the portfolio proportions $l_t = L_t/(L_t + I_t)$ and $i_t = I_t/(L_t + I_t)$ as a *two-dimensional* system of SDEs, driven by a *one-dimensional* Brownian motion.

3 Dynamics of the Portfolio Proportions

From the system of SDEs of the sub-portfolio dynamics, (1-2), the dynamics of the portfolio proportions l_t and i_t can be derived. As shown in Theorem 1, the dynamics can be described in terms of a nonlinear system of SDEs.

Theorem 1 *The dynamics of the portfolio proportions l_t and i_t can be described by a system of stochastic differential equations. This system can be stated as*

$$dl_t = a_l(l_t, i_t)dt + b_l(l_t, i_t)dB_t, \tag{3}$$
$$di_t = a_i(l_t, i_t)dt + b_i(l_t, i_t)dB_t, \tag{4}$$

where B_t is a new one-dimensional Brownian motion. The drift coefficients are

$$a_l(l_t, i_t) = -l_t\nu + i_t\delta + l_t i_t\Delta\mu - l_t^2 i_t\Delta\sigma_L^2 + l_t i_t^2\Delta\sigma_I^2, \tag{5}$$
$$a_i(l_t, i_t) = +l_t\nu - i_t\delta - l_t i_t\Delta\mu + l_t^2 i_t\Delta\sigma_L^2 - l_t i_t^2\Delta\sigma_I^2, \tag{6}$$

with $\Delta\mu = \mu_L - \mu_I$, $\Delta\sigma_L^2 = \sigma_L^2 - \sigma_L\sigma_I\rho$, *and* $\Delta\sigma_I^2 = \sigma_I^2 - \sigma_L\sigma_I\rho$. *The diffusion coefficients are*

$$b_l(l_t, i_t) = -l_t i_t \sigma, \tag{7}$$

$$b_i(l_t, i_t) = l_t i_t \sigma, \tag{8}$$

where $\sigma^2 = \Delta\sigma_L^2 + \Delta\sigma_I^2 = \sigma_I^2 + \sigma_L^2 - 2\sigma_L\sigma_I\rho$ *holds.*

PROOF: for a formal proof of this theorem, see [3], Appendix A.

This theorem provides an important result of this extended abstract. It shows that illiquidity leads to portfolio weights being stochastic processes. This is an intuitive result, as portfolio weights can no longer be fully controlled by the investor when parts of the assets in the overall portfolio cannot be traded. It is important here to stress that this result does not depend on the specification of a constant dividend yield δ and investment rate ν. This can directly be inferred from the fact that the deterministic model parameters δ and ν only appear in the drift coefficients, (5-6), but not in the diffusion coefficients, (7-8). More generally, this result will always hold as long as the portfolio position in an illiquid asset is at least temporarily irreversible.

4 Expected Portfolio Returns and Risk

In this section the impact of illiquidity on the expected portfolio returns and the portfolio variance is analyzed.

Theorem 2 *Under the developed model settings, expected instantaneous portfolio returns can be stated as*

$$\mu_P = \bar{l}\mu_L + \bar{i}\mu_I, \tag{9}$$

where $\bar{l}$ *and* $\bar{i}$ *are the expected portfolio proportions.*[2] *The instantaneous variance of portfolio returns is given by*

$$\sigma_P^2 = (\bar{l}^2\sigma_L^2 + \bar{i}^2\sigma_I^2 + 2\,\bar{l}\,\bar{i}\sigma_L\sigma_I\rho) + (\sigma_L^2 + \sigma_I^2 - 2\sigma_L\sigma_I\rho)\sigma_w^2 \tag{10}$$

where σ_w^2 *is the variance of the portfolio weights, i.e.,* $\sigma_w^2 = \sigma_{\bar{l}}^2 = \sigma_{\bar{i}}^2$. *From (10), the portfolio return variance can be decomposed into two distinct components: the standard portfolio variance,*

$$\Sigma_{SV}^2 = \bar{l}^2\sigma_L^2 + \bar{i}^2\sigma_I^2 + 2\,\bar{l}\,\bar{i}\sigma_L\sigma_I\rho, \tag{11}$$

and an additional variance term given by

[2] Note that the expectations $\bar{l}$, $\bar{i}$ and the variance σ_w^2 will, in general, be time-dependent under the developed model. However, it can be shown that the system (3-4) converges to a steady-state or long-run equilibrium as $t \to +\infty$ (for a formal proof, see [3], Appendix B). In this case $\bar{l}$, $\bar{i}$, and σ_w^2 will be constant. Here and in the following, it is assumed that such a steady-state is attained.

$$\Sigma_{AV}^2 = (\sigma_L^2 + \sigma_I^2 - 2\sigma_L\sigma_I\rho)\sigma_w^2, \tag{12}$$

that is induced by the fact that portfolio proportions are stochastic when parts of the overall portfolio are illiquid.

PROOF: for a formal proof of this theorem, see [3], Appendix C.

The results in Theorem 2 show that illiquidity does not affect expected portfolio returns. The expected return in equation (9) is equal to the standard two-asset case, except that fixed portfolio proportions are replaced here by their corresponding expectations. However, a similar result does not hold for the variance of the portfolio returns, as it can be decomposed here into two components. The first component, Σ_{SV}^2, is equal to the standard portfolio variance of the two-asset case. It corresponds to the portfolio variance an investor would have to incur in the case of fully-liquid assets and constant portfolio proportions $l = \bar{l}$ and $i = \bar{i}$. The second component, Σ_{AV}^2, is a novel contribution of this extended abstract. This additional variance is induced by the fact that portfolio proportions become stochastic when parts of the overall portfolio are illiquid. It is zero in case that $\sigma_L = \sigma_I$ and $\rho = 1$ holds. Otherwise Σ_{SP}^2 is strictly positive. Therefore, the total portfolio variance, $\Sigma_{SV}^2 + \Sigma_{AV}^2$, will always be higher than in the standard two asset reference case. This result is illustrated in Figure 1.[3] The solid line in Figure 1 represents the efficient frontier that can be constructed from a traditional mean-variance analysis in the reference case of fully liquid assets L and I and constant portfolio proportions. The dotted line represents the new efficient frontier, i.e., the efficient frontier of mean-variance combinations that arises from the results given in Theorem 2. It can be seen that the new dotted efficient frontier always lies within the space spanned by the solid traditional efficient frontier. That is, the same expected portfolio returns can only be achieved with higher level of risk when parts of the overall portfolio are illiquid. This effect is particularly apparent near the minimum standard deviations of the two efficient frontiers. Both efficient frontiers converge towards each other at their upper and lower ends. That is, the increase in portfolio risk is particularly large if the overall investment portfolio is not dominated by one of the assets.

It is also interesting to point out how the difference between the two efficient frontiers is influenced by the return correlation ρ. The magnitude of the difference between the two efficient frontiers depends on the additional variance Σ_{AV}^2. Interestingly, this term increases when the correlation between the assets returns decreases. It is a main result from the traditional mean-variance theory that portfolio risk decreases with decreasing levels of the asset return correlations. This result still continues to hold under the developed model setting because of the standard variance term given by equation (11). However, it is extended here. The lower the return correlation ρ, the higher is the additional variance of portfolio returns caused by illiquidity. Stated differently, illiquidity particularly increases portfolio risk if the return correlation between the liquid and illiquid assets is low.

[3] The parameter constellation used to create the figure is as follows: $\delta = 0.3$, $\mu_I = 0.30$, $\sigma_I = 0.3$, $\mu_L = 0.15$, $\sigma_L = 0.15$ and $\rho = -0.75$.

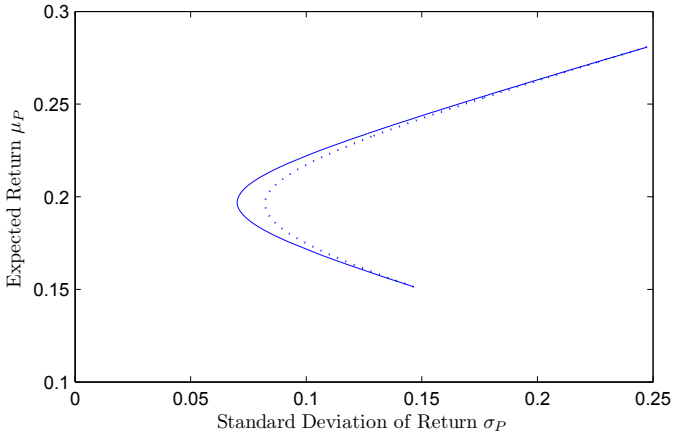


Fig. 1 Efficient Frontier of Traditional Mean-Variance Analysis (Solid Line) Compared to the New Efficient Frontier (Dotted Line)

5 Conclusion

This extended abstract makes two key points about illiquidity and asset pricing. First, it shows that illiquidity leads to portfolio proportions being stochastic. Second, it shows that stochastic proportions result in an additional variance that increases overall portfolio risk. This additional variance is inversely related to the return correlation between the illiquid and liquid asset. In general, it is often argued in the literature that adding illiquid alternative assets to a portfolio is particularly useful because of the low return correlation of these assets with traditional traded assets. The results presented here show that parts of the diversification benefits attributed to illiquid alternative assets can be offset by the increase in portfolio variance caused by illiquidity. Therefore, the results imply that the benefits of adding illiquid assets to a portfolio could be substantially lower than typically perceived.

References

1. Viral V. Acharya and Lasse Heje Pedersen. Asset pricing with liquidity risk. *Journal of Financial Economics*, 77(2): 375–410, 2005.
2. Yakov Amihud and Haim Mendelson. Liquidity, maturity, and the yields on U.S. treasury securities. *Journal of Finance*, 46(4): 1411–1425, 1991.
3. Axel Buchner. New insights on asset pricing and illiquidity. Working paper, Technical University of Munich, 2010.
4. Francis A. Longstaff. Optimal portfolio choice and the valuation of illiquid securities. *Review of Financial Studies*, 14(2): 407–431, 2001.
5. Lubos Pastor and Robert F. Stambaugh. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3): 642–685, 2003.

Robust Risk Management in Hydro-Electric Pumped Storage Plants

Apostolos Fertis and Lukas Abegg

Abstract The hydro-electric pumped storage plant management can be formulated as a multi-stage stochastic optimization problem. Recent research has identified efficient policies that minimize the risk of the plant value in a finite horizon framework by aggregating the spot price information into epochs. The most widely accepted models describe the spot price as being affected by a number of factors, each one following a mean-reverting type stochastic process. Yet, the conditions that affect those factors are rather volatile, making the adoption of a single probability model dangerous. Indeed, historical data show that the estimation of the stochastic models parameters is heavily dependent on the set of samples used. To tackle this problem, we follow a robust risk management approach and propose a method to compute optimal dispatch policies that minimize the Robust CVaR of the final plant value. We compare the optimal-Robust CVaR to the optimal-CVaR policies in probability models constructed using spot price data from EPEX for the period between 2005 and 2010. The computational overhead of computing optimal-Robust CVaR policies is linear in the number of scenarios and the computational complexity can be adjusted to achieve the desired level of accuracy in planning.

1 Introduction

Many hydro-electric pumped storage plants are equipped with two water reservoirs at different levels. They can produce electricity using the potential energy of the water in the upper reservoir and they can store electricity by pumping water from

Apostolos Fertis
Institute for Operations Research (IFOR), Eidgenössische Technische Hochschule Zürich (ETH Zurich), HG G 21.3, Rämistrasse 101, 8092 Zürich, Switzerland e-mail: afertis@ifor.math.ethz.ch

Lukas Abegg
Department of Mathematics (D-MATH), Eidgenössische Technische Hochschule Zürich (ETH Zurich), Rämistrasse 101, 8092 Zürich, Switzerland e-mail: labegg@student.ethz.ch

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_16,

the lower to the upper reservoir. They are connected to an electricity exchange, being able to buy or sell electricity at the spot price. They would like to benefit by buying electricity at low prices and selling it in high prices. This description shows that the hydro-electric plant trading strategy constitutes a multistage stochastic programming problem, where the objective is to maximize the final value of the plant. The uncertainty in the evolution of the electricity market makes necessary the consideration of risk.

In quantitative risk management, risk measures are functionals defined on a space of random variables describing the uncertain future value of a financial position. Recent research has identified certain classes of risk measures with desirable properties. Artzner, Delbaen et al. defined the class of coherent risk measures, and proved that any coherent risk measure can be expressed as the worst-case expectation of the random return when the probability measure used in the expectation varies in some set [1]. Föllmer and Schied defined the class of convex risk measures, a superclass of the class of coherent risk measures, and proved that any convex risk measure can be expressed as the conjugate function of some "penalty" function defined in the dual space of the random variables space [6].

Risk evaluation is based on the probability model assumed for the financial position in consideration. But the models do not usually fully represent reality. Also, the estimation of the model parameters is affected by the sample data which capture only some snapshots of the market evolution. In many cases, there exist several candidate models, which are different scenario probability distributions, and it can be assumed that the actual model is a mixture of those. Nevertheless, it is difficult to determine the probability distribution over the scenarios, but it might be safer to determine a certain range for it. To overcome this difficulty and take into consideration all risk sources hidden in the scenarios, the robust risk approach has been proposed, as, for example, in [5, 8]. In these approaches, a robust optimization strategy is followed to tackle the uncertainty in the underlying probability measure.

Several models for the electricity spot prices have been proposed [2, 3]. Most of the models assume that the spot price is a weighted sum of a seasonality term and usually three mean-reverting processes with different mean-reverting rates. To reduce the complexity in calculating optimal-risk policies for the hydro-electric power plants Densing proposed the idea to aggregate the spot price information into epochs and model the occupation times of price levels as a weighted sum of mean-reverting processes [4]. Densing computes policies that minimize the expected value of the final value of the plant with the multi-period CVaR being restricted to an upper limit.

In this paper, we use Densing's idea of occupation times to compute policies that minimize the RCVaR of the final value of the plant, and compare them to the policies that minimize the CVaR. In particular, we construct a discrete probability model which represents the occupation times as a mean-reverting process. We propose an estimator for the parameters of this model which returns a model with a fixed number of states by applying the k-means clustering algorithm. We express the linear optimization problems that calculate the optimal-CVaR and optimal-RCVaR policies.

We compute the optimal-CVaR and optimal-RCVaR policies using data from EPEX for the period between 2005 and 2010. We compare the performance of the two policies in the presence of uncertainties in the model.

2 Spot Price Model

2.1 Description of the Model

The spot price information is aggregated into epochs. Instead of modeling the evolution of the spot price, we model the evolution of the occupation times of different price levels in the epochs. We define a set of m price levels denoted by $\bar{s}_j$, $j = 1, 2, \ldots, m$, and $\bar{s}_0 = \bar{s}_{m+1} = \infty$. Each price level $\bar{s}_j$ is the center of the price interval I_j, namely $I_j = ((\bar{s}_j - \bar{s}_{j-1})/2, (\bar{s}_{j+1} - \bar{s}_j)/2]$, $j = 1, 2, \ldots, m$. An epoch consists of H hours. The spot price at the hour i of epoch t is denoted by $s_{t,i}$, $i = 1, 2, \ldots, H$. If we denote by $\mathbb{I}$ the indicator function, the occupation time of level j during epoch t is then given by $F_t^j = (1/H) \sum_{i=1}^{H} \mathbb{I}_{s_{t,i} \in I_j}$, $j = 1, 2, \ldots, m$. Vector $\boldsymbol{F}_t$ contains the occupation times F_t^j in its entries. The occupation times are assumed to follow a mean-reverting process. Since occupation times should satisfy $F_t^j \geq 0$, $j = 1, 2, \ldots, m$, and $\sum_{j=1}^{m} F_t^j = 1$, to express their evolution, we define the projection operator $\Pi(\boldsymbol{F}) = (1/\sum_{j=1}^{m}(F^j)^+)\,[(F^1)^+, (F^2)^+, \ldots, (F^m)^+]^T$, $\boldsymbol{F} \neq \boldsymbol{0}$, $\Pi(\boldsymbol{0}) = [1/m, 1/m, \ldots, 1/m]^T$. Considering ε_t to be independent random variables that follow the same distribution $\mathcal{E}$ and using operator Π, we can describe the evolution of the occupation times by

$$\boldsymbol{F}_t = \Pi(\boldsymbol{F}_{t-1} + \theta(\boldsymbol{\mu} - \boldsymbol{F}_{t-1}) + \boldsymbol{\varepsilon}_t). \tag{1}$$

The model parameters are θ, $\boldsymbol{\mu}$ and the probability distribution of $\mathcal{E}$.

2.2 Estimation of the Model

In this section, we propose an estimator for the occupation times model described in Eq. (1). Using spot price data for a particular time range and dividing this time range into epochs, we can obtain sample data $\boldsymbol{F}_t$, $t = 1, 2, \ldots, T$, for the occupation times. The estimate for $\boldsymbol{\mu}$ is then $\hat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \boldsymbol{F}_t/T$. Parameter θ is estimated by a linear regression between $F_t^j - F_{t-1}^j$ and $\hat{\mu}_j - F_{t-1}^j$, $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, T$. Having estimated $\boldsymbol{\mu}$ and θ, we can calculate the error terms $\varepsilon_t \in \mathbb{R}^m$, $t = 1, 2, \ldots, T$, on which we apply the k-means clustering method with $k = q$. We use the size and the centers of the q clusters to define a q-state discrete probability distribution for ε_t.

3 Optimal-Risk Policies

In this section, we describe the computation of optimal-CVaR and optimal-RCVaR policies for the final value of the plant in a finite horizon scheme. For any discrete probability distribution of the error terms ε_t, we can construct a scenario tree, where each node at level t corresponds to a different realization of random variables ε_1, ε_2, …, ε_t. If T is the length of the planning horizon, then, the depth of this tree is T. Set $\mathcal{N}_t$ contains the nodes at level t and set $\mathcal{N} = \cup_{t=1,2,\ldots,T}\mathcal{N}_t$. Using Eq. (1), we can compute the occupation times vector $\boldsymbol{F}_{t,i}$ for any node $i \in \mathcal{N}_t$ at any level $t = 1, 2, \ldots, T$.

The plant policy is described by $u_{t,i}^{+,j}$, $u_{t,i}^{-,j}$, $i \in \mathcal{N}_t$, $i \in \mathcal{N}_t$, $t = 1, 2, \ldots, T$, where $u_{t,i}^{+,j}$, $u_{t,i}^{-,j}$ is the amount of energy to bid for selling or buying respectively at price level j, $j = 1, 2, \ldots, m$. By the rules set by energy exchanges, it should hold that $u_{t,i}^{+,j} \leq u_{t,i}^{+,j+1}$ and $u_{t,i}^{+,j} \geq u_{t,i}^{+,j+1}$. If c is the loss factor describing the loss in storing the energy and I is the assumed constant inflow of the water during an epoch, the dynamics that describe the evolution of the water level and the cash held are

$$L_{t,i_1} = L_{t,i_0} + I + H \sum_{j=1}^{m} (-F_{t,i_1}^{j} u_{t-1,i_0}^{+,j} + cF_{t,i_1}^{j} u_{t-1,i_0}^{-,j}), \tag{2}$$

$$C_{t,i_1} = C_{t,i_0} + H \sum_{j=1}^{m} (F_{t,i_1}^{j} u_{t-1,i_0}^{+,j} \bar{s}_j - F_{t,i_1}^{j} u_{t-1,i_0}^{-,j} \bar{s}_j), \tag{3}$$

where node i_1 at level t is a child of node i_0 at level $t-1$. Since the scenario tree contains information only about the occupation times and not the actual spot prices, the value $V_{t,i}$ of the plant at node i in level t is approximated by $V_{t,i} = C_{t,i} + \sum_{j=1}^{m} F_{t,i}^{j} L_{t,i} \bar{s}_j$. Since the dynamics are described by linear inequalities, we can compute optimal-CVaR policies by solving a linear optimization problem with $O(q^T)$ constraints [7].

In the case that the actual model is a mixture of a set of different models, we construct a scenario tree for each scenario model and by choosing a box uncertainty set for the probability distribution over the scenarios, we can compute optimal-RCVaR policies, as in [5]. The number of constraints in the resulting linear optimization problem is $O(rq^T)$, where r is the number of scenario probability distributions considered.

4 Numerical Results

We used the EPEX electricity prices of each year between 2005 and 2010 to estimate the parameters of our model. In this way, we constructed 5 different models for the occupation times evolution. Using the methodology described in Section 3, we computed the optimal-CVaR policy assuming that the actual model is an equally

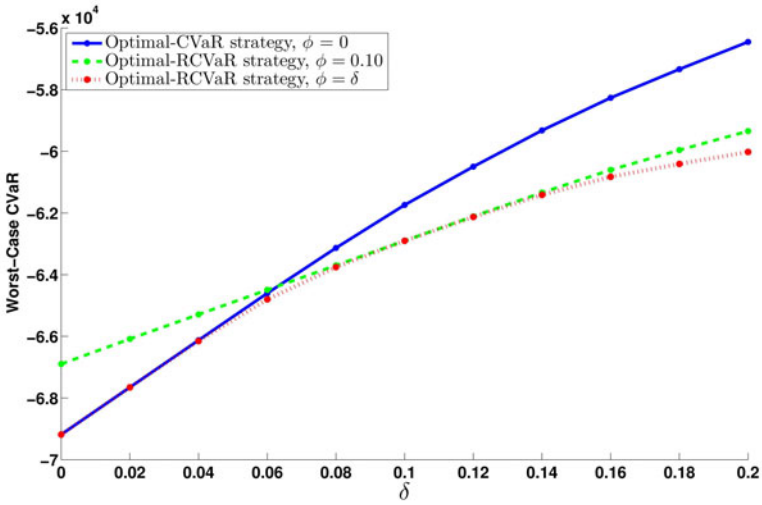


Fig. 1 Comparison of optimal-CVaR and optimal-Robust CVaR with $\phi = 0.10$ policies.

weighted sum of the 5 different models. We, also, compute the optimal-CVaR policy for different values of ϕ, where ϕ is the size of the box uncertainty set for the actual probability distribution. The nominal probability distribution is assumed to be the equally weighted sum of the 5 different models.

The policies are evaluated by assessing the worst-case CVaR of the final value of the plant, when the actual probability distribution resides in a δ-sized box uncertainty set around the nominal distribution. In Figure 1, we see the performance of the different policies. We observe that if we choose the protection level $\phi = 0.10$, for a quite big range of the large values of the real uncertainty δ, the optimal-RCVaR policy achieves a substantially lower risk than the optimal-CVaR policy, and it is quite near the ideal robust policy with $\phi = \delta$. In Figure 2, we see the same results in the case that we add a maximum water level constraint $L \leq 1500$ MWh. In this case, the differences among the policies are smaller, implying the maximum water level constraint pushes towards a more robust policy.

5 Conclusions

The robust risk management approach offers a new way to deal with uncertainty in the electricity price models, when computing hydro-electric plant policies. Although it suffers from the curse of dimensionality in multistage stochastic programming, as the optimal-CVaR approach does, it can be used to overcome model uncertainty and take safer decisions.

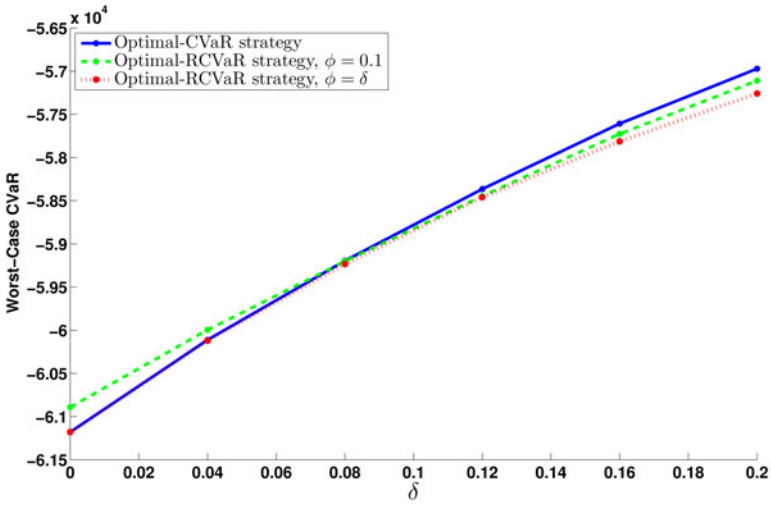


Fig. 2 Comparison of optimal-CVaR and optimal-Robust CVaR with $\phi = 0.10$ policies with the maximum water level restriction $L \leq 1500$ MWh.

References

1. P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent Risk Measures. *Mathematical Finance*, 9(3): 203–228, 1999.
2. M. T. Barlow, Y. Gusev, and M. Lai. Calibration of multifactor models in electricity markets. *International Journal of Theoretical and Applied Finance*, 7(2): 101–120, 2004.
3. O. E. Barndorff-Nielsen and S. Shepard. Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 167–241, 2002.
4. M. Densing. *Hydro-Electric Power Plant Dispatch-Planning – Multi-Stage Stochastic Programming with Time-Consistent Constraints on Risk*. PhD Thesis, ETH Zurich, 2007.
5. A. Fertis, M. Baes, and H.-J. Lüthi. Robust Risk Management. *Submitted for publication*, 2010.
6. H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4): 429–447, 2002.
7. A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming, Modeling and Theory*. Society for Industrial and Applied Mathematics and Mathematical Programming Society, Philadephia, PA, USA, 2009.
8. S. Zhu and M. Fukushima. Worst-Case Conditional Value-at-Risk with Application to Robust Portfolio Management. *Operation Research*, 57(5): 1155–1168, September-October 2009.

Copulas, Goodness-of-Fit Tests and Measurement of Stochastic Dependencies Before and During the Financial Crisis

Peter Grundke and Simone Dieckmann

Abstract By means of goodness-of-fit tests, it is analyzed for a multitude of portfolios consisting of different asset classes whether the stochastic dependence between the portfolios' constituents can be adequately described by multivariate versions of some standard parametric copula functions. Furthermore, it is tested whether the stochastic dependence between the returns of different asset classes has changed during the recent financial crisis.

1 Introduction

The knowledge of the multivariate stochastic dependence between random variables is of crucial importance for many finance applications. For example, the comovement of world equity markets is frequently interpreted as a sign of global economic and financial integration. Portfolio managers have to know the stochastic dependence between different asset classes to exploit diversification effects between these asset classes by adequate asset allocation. Market risk managers have to know the stochastic dependence between the market risk-sensitive financial instruments in their portfolio to compute risk measures, such as value-at-risk or expected shortfall. Credit risk managers have to know the stochastic dependence between latent variables driving the credit quality of their obligors to compute risk measures for their credit portfolios. Financial engineers constructing multi-asset financial derivatives have to know the stochastic dependence between the different underlyings to correctly price the derivatives.

Peter Grundke
University of Osnabrück, Chair of Banking and Finance, Katharinenstraße 7, 49069 Osnabrück, Germany, e-mail: peter.grundke@uni-osnabrueck.de

Simone Dieckmann
University of Osnabrück, Chair of Banking and Finance, Katharinenstraße 7, 49069 Osnabrück, Germany, e-mail: simone.dieckmann@uni-osnabrueck.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_17,

Copulas are a popular concept for measuring the stochastic dependence between random variables. They describe the way that the marginal distributions of the components of a random vector have to be coupled to yield the joint distribution function. However, the parameters of an assumed parametric family of copulas have to be estimated and, afterwards, the adequacy of the copula assumption has to be proved. One way of doing this adequacy check is by performing goodness-of-fit (gof) tests for copulas. We employ gof tests based on the Rosenblatt transformation and based on the empirical copula for testing various copula assumptions with respect to the stochastic dependence between the returns of different asset classes (credit, stocks, commodities, real estate and so on). The two main questions we want to answer are:

1. Is it possible to describe the multivariate stochastic dependence between the returns of different asset classes by some standard parametric copula?
2. If so, how did the stochastic dependence between the returns of different asset classes change during the recent financial crisis? A satisfying answer to this second question could be of importance for multivariate stress tests based on historical scenarios.

The remainder of the paper is structured as follows: In Section 2, a short review of the methodology is given. In Section 3, the data and its processing and filtering are described. Section 4 contains the main empirical results. Finally, in Section 5, conclusions are shortly sketched.

2 Methodology

Any joint distribution function of random variables contains information on the marginal behavior of the single random variables *and* on the dependence structure of all random variables. Copula functions only deal with the stochastic dependence and provide a way of isolating the description of the dependence structure from the joint distribution function. According to Sklar's Theorem, any joint cumulative density function (cdf) $F_{X_1,...,X_D}(x_1,...,x_D)$ of D random variables $X_1,...,X_D$ can be written in terms of a copula function $C(u_1,...,u_D)$ and the marginal cdf's $F_{X_1}(x_1),...,F_{X_D}(x_D)$:[1]

$$F_{X_1,...,X_D}(x_1,...,x_D) = C(F_{X_1}(x_1),...,F_{X_D}(x_D)). \tag{1}$$

For recovering the copula function of a multivariate distribution $F_{X_1,...,X_D}(x_1,...,x_D)$, the method of inversion can be applied:

$$C(u_1,...,u_D) = F_{X_1,...,X_D}(F_{X_1}^{-1}(u_1),...,F_{X_D}^{-1}(u_D)) \tag{2}$$

where $F_{X_1}^{-1}(\cdot),...,F_{X_D}^{-1}(\cdot)$ are the inverse marginal cdf's. For continuous random variables $X_1,...,X_D$, the copula function is unique. In our context, $F_{X_1}(x_1),...,F_{X_D}(x_D)$

[1] Standard references for copulas are [11] and [13].

are the marginal cdf's of the daily log-returns of indices that represent specific asset classes.

In this paper, we employ the following parametric copulas: Gaussian copula, t-copula, Gumbel copula, Clayton copula, and generalized Clayton Copula. As the parametric families of copulas that we test belong to different classes of copulas (elliptical and Archimedean) each with a different extent of tail dependence (without and with (symmetric or asymmetric) tail dependence), we believe that we test a reasonable coverage of parametric families of copulas. This is why we restrict ourselves to (low-) parametric families of copulas instead of estimating non-parametric (empirical) copulas.

For estimating a multivariate model consisting of the marginal distributions and the copula, several approaches exist (see [11] or [4] for an overview). One possibility for estimating the parameters of the copula function is the canonical maximum likelihood estimation (also called pseudo-maximum likelihood). For this method, there is no need to specify the parametric form of the marginal distributions because these are replaced by the empirical marginal distributions. Thus, only the parameters of the copula function have to be estimated by maximum likelihood (see [4], p. 160). We employ this approach, however, partly in a modified version to further reduce the computational burden. For example, for estimating the parameters of a t-copula, first, the correlation parameters are computed by the method-of-moments inverting the relationship $\rho_\tau(X_i,X_j) = (2/\pi)\arcsin\rho_{ij}$ between Kendall's tau and the correlation parameters of the t-copula (see [12], p. 231). Afterwards, the remaining degree of freedom is estimated by maximum (pseudo-) likelihood. Similarly, the correlation parameters of a Gaussian copula can be computed using their relationship $\rho_S(X_i,X_j) = (6/\pi)\arcsin 0.5\rho_{ij}$ to Spearman's rho.

From the many gof tests proposed in the literature in the recent years, we choose one test based on the Rosenblatt transformation combined with the Anderson Darling and the Cramér-von Mises test statistic, respectively (see, e.g., [3, 6]) and one test based on the empirical copula process combined with the Cramér-von Mises test statistic (see [7]). Both tests exhibit a relatively good performance in power comparisons of different gof tests (see [2, 8]). Unfortunately, for both test, the empirical distribution of the test statistics under the null hypothesis has to be computed by bootstrapping due to the fact that the true marginal distributions as well as the true parameters of the copulas are unknown (see the above quoted papers for details concerning this procedure).[2]

3 Data

In total, we consider six different asset classes:

[2] For all gof tests and all copulas, the number of bootstrap simulations is set equal to 1,000.

- credit (represented by the DJ iTraxx Credit Default Swap (CDS) master index for Europe and six sector-specific subindices),[3]
- stocks (represented by six major stock indices: MSCI World, S&P 500, DAX 30, DJ Euro STOXX 50, FTSE 100, CAC 40),
- bonds (represented by the Datastream EMU Benchmark 10 year DC Govt. index),
- currencies (represented by the USD/EUR exchange rate),
- commodities (represented by the Reuters/Jefferies CRB index),
- real estate (represented by the FTSE EPRA/Nareit Global index).[4]

The daily data for all asset classes is gathered from Datastream and Bloomberg and covers the period January 2, 2006 to October 21, 2008. Based on daily midpoint quotes, log-returns are computed. Each time series is divided into two subsamples covering the time periods January 2, 2006 to June 29, 2007 and July 2, 2007 to October 21, 2008, respectively. The first sample represents the pre-crisis sample whereas the second sample is considered to be the crisis sample. The length of the data sample is shortened because the time series are adjusted for holidays and filtered by ARMA-GARCH models (see below). In detail, this means that twenty holidays in the pre-crisis and eighteen holidays in the crisis sample were excluded. The results of the ARMA-GARCH filter show different lags of the AR process element. To get time series with equal length, the first values of those time series with a lower lag are deleted. In consequence, the pre-crisis sample and the crisis sample contain N=365 and N=321 values, respectively.

For three different groups of portfolios, the null hypotheses that specific parametric copulas adequately describe the stochastic dependence between the portfolios' constituents are tested. The first group of portfolios consists of all possible two- to six-dimensional combinations of the six sector-specific subindices of the DJ iTraxx CDS master index for Europe ("DJ iTraxx portfolios"). The second group of portfolios ("stock portfolios") are formed by all two- to five-dimensional combinations of the five major stock indices (S&P 500, DAX 30, DJ Euro STOXX 50, FTSE 100, CAC 40). The third group of portfolios ("mixed portfolios") consist of all possible two- to six-dimensional combinations of the following asset classes: credit (DJ iTraxx CDS master index), stocks (MSCI World), bonds (EMU index), currencies (USD/EUR exchange rate), commodities (CRB index), and real estate (FTSE EPRA/Nareit index).

The gof tests need independent and identically distributed (i.i.d.) data as input. However, plots of the autocorrelation function and the partial autocorrelation function of the returns and the squared returns show that in general, as expected, they exhibit autocorrelation and time-varying conditional volatilities. These visual results are also confirmed by formal statistical tests, such as the Ljung-Box test which rejects the null hypothesis of no autocorrelation and Engle's Lagrange multiplier (LM) test which indicates that there are indeed ARCH effects. To remove autocorrelation and conditional heteroscedasticity in the univariate time series of returns, an

[3] See [9] for details.

[4] All indices are price indices.

ARMA model with GARCH errors is fitted to the raw returns of each subindex and asset class, respectively.[5] Finally, the gof tests for copulas are applied to the filtered returns[6]

4 Main Results

Several main results can be stated: First, whether a specific copula assumption can be rejected or not, crucially depends on the asset class and the time period considered. For example, the multivariate t-copula is predominantly rejected in the crisis sample for the higher-dimensional DJ iTraxx portfolios, but it is predominantly not rejected in the pre-crisis period and it is also predominantly not rejected for stock portfolios (in the pre-crisis period) and for mixed portfolios (in both periods). Second, different gof tests for copulas can yield very different results and these differences can be different for different asset classes and for different tested copulas. For example, employing the gof test based on the empirical copula, the number of rejections of the Gaussian copula in the pre-crisis sample decreases dramatically for DJ iTraxx portfolios (compared to the rejection rates that the gof test based on the Rosenblatt transformation yields). However, for the null hypothesis of a t-copula for DJ iTraxx portfolios in the crisis period, the number of rejections is highest when the gof test based on the empirical copula is used. In consequence, these findings raise some doubt with respect to the reliability of the results based on gof tests for copulas, especially when applied to relatively short time series. Third, even when using different gof tests for copulas, it is not always possible to differentiate between various copula assumptions. For example, for stock portfolios in the time period before the crisis, neither the null hypothesis of a Gaussian copula nor the null hypothesis of t-copula with a relatively low average degree of freedom can be rejected. With respect to the second question raised in the introduction, whether risk dependencies changed during the recent financial crisis, it can be stated that there is a tendency for increased rejection rates during the crisis. This is true across different portfolios, different copulas in the null hypothesis and different test statistics. Thus, risk dependencies are seemingly more complex during the crisis and tend to be not adequately described by standard parametric copulas. As expected, correlations and the degree of freedom of the t-copula increase on average (beside for the stocks portfolios). Both results imply a larger degree of tail dependence during the crisis than before the crisis. However, the results also raise some concerns over the suitability of gof tests for copulas as a diagnostic tool for identifying stressed risk dependencies, because the typical length of time series in stress scenarios might be to short for these kind of tests.

[5] See [12], p. 148.

[6] See [9] for details.

5 Conclusions

Finally, we think that it would be promising in future research to consider also some more advanced copula concepts (such as the grouped t-copula, the skewed t-copula (for both, see [5]), conditional copulas (see, e.g., [14]), mixed copulas (see, e.g., [10]), or pair-copula constructions (see, e.g., [1])) for modelling the multivariate stochastic dependence between returns of different asset classes, in particular in turbulent times on financial markets.

Acknowledgements We thank Daniel Berg for kindly providing his R-package copulaGOF.

References

1. K. Aas, C. Czado, A. Frigessi, and H. Bakken. "Pair-copula constructions of multiple dependence". *Insurance: Mathematics and Economics*, 44(2): 182–198, 2009.
2. D. Berg. "Copula goodness-of-fit testing: an overview and power comparison". *European Journal of Finance*, 15(7&8): 675–701, 2009.
3. W. Breymann, A. Dias, and P. Embrechts. "Dependence structures for multivariate high-frequency data in finance". *Quantitative Finance*, 3: 1–14, 2003.
4. U. Cherubini, E. Luciano, and W. Vecciato. *"Copula methods in finance"*. Wiley, Chichester, 2004.
5. S. Demarta and A.J. McNeil. The t copula and related copulas. *"International Statistical Review"*, 73(1): 111–129, 2005.
6. J. Dobrić and F. Schmid. "A goodness of fit test for copulas based on Rosenblatt's transformation". *Computational Statistics & Data Analysis*, 51(9): 4633–4642, 2007.
7. C. Genest and B. Rémillard. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques*, 44(6): 1096–1127, 2008.
8. C. Genest, B. Rémillard, and D. Beaudoin. "Goodness-of-fit tests for copulas: A review and a power study". *Insurance: Mathematics and Ecnomics*, 44: 199–214, 2009
9. P. Grundke. "Changing default risk dependencies during the subprime crisis: DJ iTraxx subindices and goodness-of-fit-testing for copulas". *Review of Managerial Science*, 4(2): 91–118, 2010.
10. L. Hu. "Dependence patterns across financial markets: a mixed copula approach". *Applied Financial Economics*, 16: 717–729, 2006.
11. H. Joe. *"Multivariate models and dependence concepts"*. Chapman & Hall, London, 1997.
12. A.J. McNeil, R. Frey, and P. Embrechts. *"Quantitative Risk Management"*. Princeton University Press, Princeton and Oxford, 2005.
13. R.B. Nelson. *"An introduction to copulas. (2nd ed.)"*. Springer, New York, 2006.
14. A. Patton. "Modelling asymmetric exchange rate dependence". *International Economic Review*, 47(2): 527–556, 2006.

Confidence Intervals for Asset Correlations in the Asymptotic Single Risk Factor Model

Steffi Höse and Stefan Huschens

Abstract The asymptotic single risk factor (ASRF) model, which has become a standard credit portfolio model in the banking industry, is parameterized by default probabilities and asset (return) correlations. In this model, individual and simultaneous confidence intervals for asset correlations are developed on the basis of observed default rates. Since the length of these confidence intervals depends on the confidence level chosen, they can be used to define stress scenarios for asset correlations.

1 Introduction

Financial institutions are under pressure to implement meaningful sensitivity analyses and sound stress testing practices for the parameters of the credit portfolio models used. Although these models differ in their assumptions and complexity, the typical input parameters are default probabilities and asset correlations. As the literature focuses mainly on the point estimation of these parameters [5, 6, 13] and on the interval estimation of default probabilities [3, 8, 10, 14], this paper concentrates on the statistical interval estimation of asset correlations in the ASRF model of Basel II. One approach used to estimate implicit asset correlations in this model is based on the means and variances of default rates, see [2, p. 117]. In this paper, an alternative approach based on transformed default rates is used to derive point and interval estimators for asset correlations.

Steffi Höse
Technische Universität Dresden, e-mail: steffi.hoese@tu-dresden.de

Stefan Huschens
Technische Universität Dresden, e-mail: stefan.huschens@tu-dresden.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_18,

2 Statistical Model

A credit portfolio classified into R risk categories (e. g. rating classes, segments, industry sectors, risk buckets) is considered for T subsequent periods. By assuming an ASRF model with infinite granularity in each risk category, as suggested by [1, p. 64], the distribution of the credit portfolio loss in period $t \in \{1,\dots,T\}$ can be approximated by the distribution of the systematic part of the portfolio loss $\sum_{r=1}^{R} w_{tr}\tilde{\pi}_{tr}$. Thereby, the loss contributions $w_{tr} \geq 0$ of all loans assigned to risk category r in period t are assumed to be known and the stochastic default probabilities

$$\tilde{\pi}_{tr} \stackrel{\text{def}}{=} \Phi\left(\frac{\Phi^{-1}(\pi_r) - \sqrt{\rho_r}Z_t}{\sqrt{1-\rho_r}}\right), \qquad Z_t \overset{\text{i.i.d.}}{\sim} N(0,1) \tag{1}$$

of all obligors in risk category r and period t are functions of the model parameters $0 < \pi_r < 1$ and $0 < \rho_r < 1$.[1] Here, Φ denotes the distribution function of the standardized Gaussian distribution and Φ^{-1} its inverse, which is also known as the probit function. The stochastic default probability $\tilde{\pi}_{tr}$ depends monotonically on the random variable Z_t, which represents the systematic risk factor of period t. This risk factor simultaneously affects the creditworthiness of the obligors in all risk categories. Equation (1) results from a single risk factor model, in which the stochastic default probability $\tilde{\pi}_{tr}$ is the conditional default probability of all obligors in category r, given the systematic risk factor Z_t. It can be shown that $\tilde{\pi}_{tr}$ follows a Vasicek distribution with parameters π_r and ρ_r [17], where $\pi_r = \mathbb{E}[\tilde{\pi}_{tr}]$ is the *mean* or *unconditional default probability* of all obligors in category r [2, p. 90] and ρ_r can be interpreted as the *asset (return) correlation* of two different obligors in the same category r. Thus, the ASRF model is parameterized by the vector $\boldsymbol{\pi} = (\pi_1,\dots,\pi_R) \in]0,1[^R$ of default probabilities and the vector $\boldsymbol{\rho} = (\rho_1,\dots,\rho_R) \in]0,1[^R$ of asset correlations.

Within the ASRF model, the stochastic default probabilities $\tilde{\pi}_{tr}$ can be identified with observed default rates by assuming *infinite granularity* for each risk category and ignoring sampling errors caused by the finite number of obligors in each category. By using this idea, the observable *probits*[2]

$$Y_{tr} \stackrel{\text{def}}{=} \Phi^{-1}(\tilde{\pi}_{tr}) = \frac{\Phi^{-1}(\pi_r) - \sqrt{\rho_r}Z_t}{\sqrt{1-\rho_r}} \tag{2}$$

of the stochastic default probabilities $\tilde{\pi}_{tr}$ are linear transformations of the risk factors $Z_t \sim N(0,1)$. Therefore the probits are Gaussian distributed with parameters

$$\mu_r \stackrel{\text{def}}{=} \mathbb{E}\big[Y_{tr}\big] = \frac{\Phi^{-1}(\pi_r)}{\sqrt{1-\rho_r}}, \qquad \sigma_r^2 \stackrel{\text{def}}{=} \mathbb{V}\big[Y_{tr}\big] = \frac{\rho_r}{1-\rho_r} \tag{3}$$

and with covariances $\mathbb{C}ov\big[Y_{tr}, Y_{ts}\big] = \sigma_r\sigma_s$ for $r,s = 1,\dots,R$, $r \neq s$ and $t = 1,\dots,T$.

[1] For the ASRF model compare [7], [16, pp. 309–312], [17] and [18].

[2] [9, pp. 194ff.] derives simultaneous confidence sets based on probits of default rates and called this approach the transformation method.

3 Confidence Intervals for Asset Correlations

For the random variables Y_{tr}, a parametric model with parameters $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is given by (1) and (2). An alternative parameterization is $(\mu_1, \ldots, \mu_R, \sigma_1^2, \ldots, \sigma_R^2)$, since the equations in (3) define bijections between $(\pi_r, \rho_r) \in]0,1[^2$ and $(\mu_r, \sigma_r^2) \in \mathbb{R} \times]0, \infty[$ for all $r = 1, \ldots, R$. For $T > 1$, the variance σ_r^2 in (3) may be estimated by the sample variance

$$\hat{\sigma}_r^2 \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{T} (Y_{tr} - \bar{Y}_r)^2 \quad \text{with} \quad \bar{Y}_r \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^{T} Y_{tr}.$$

By using $\rho_r = \sigma_r^2/(1+\sigma_r^2)$, which is equivalent to the second equation in (3), the estimator $\hat{\sigma}_r^2$ for the variance σ_r^2 can be transformed into the estimator

$$\hat{\rho}_r = \frac{\hat{\sigma}_r^2}{1+\hat{\sigma}_r^2} \tag{4}$$

for the asset correlation ρ_r. Since $(\bar{Y}_1, \ldots, \bar{Y}_R, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_R^2)$ is the maximum likelihood estimator for $(\mu_1, \ldots, \mu_R, \sigma_1^2, \ldots, \sigma_R^2)$, the vector $(\bar{Y}_1, \ldots, \bar{Y}_R, \hat{\rho}_1, \ldots, \hat{\rho}_R)$ is the maximum likelihood estimator for $(\mu_1, \ldots, \mu_R, \rho_1, \ldots, \rho_R)$. For a direct derivation of the maximum-likelihood estimator $\hat{\rho}_r$ see [4, p. 8].

Confidence intervals for the parameter σ_r^2 can be constructed by using standard textbook methods based on the point estimator $\hat{\sigma}_r^2$ for σ_r^2 and on the chi-squared distribution, since the random vectors $(Y_{t1}, \ldots, Y_{tR})$ are stochastically independent for $t = 1, \ldots, T$. By using (4), confidence intervals for σ_r^2 can be transformed into confidence intervals for the asset correlations ρ_r.[3]

Theorem 1 *Let the ASRF model given by (1) hold for $T > 1$ periods. Let $0 < \alpha < 1$ and let $\chi^2_{T-1,p}$ be the p-quantile of a chi-squared distribution with $T-1$ degrees of freedom. Then for $r = 1, \ldots, R$ each stochastic interval*

$$I_r = \left[\frac{T\hat{\rho}_r}{T\hat{\rho}_r + \chi^2_{T-1,1-\alpha/2}(1-\hat{\rho}_r)}, \frac{T\hat{\rho}_r}{T\hat{\rho}_r + \chi^2_{T-1,\alpha/2}(1-\hat{\rho}_r)} \right]$$

is a $(1-\alpha)$-confidence interval for the asset correlation ρ_r, i. e.

$$\Pr(\rho_r \in I_r) = 1 - \alpha \quad \textit{for all } (\boldsymbol{\pi}, \boldsymbol{\rho}).$$

The confidence intervals I_r have the constant coverage probability $(1-\alpha)$. Since the quantile $\chi^2_{T-1,p}$ is strictly increasing in p, the resulting length of I_r is strictly decreasing in α. This monotonicity facilitates the use of confidence intervals for the purpose of sensitivity analysis. Decreasing α towards zero leads to increasing sets of stress scenarios for the asset correlation parameter. However, it should be noted that a very small α leads to a useless confidence interval I_r.

[3] For the confidence intervals given in Theorem 1 compare [9, Theorem 6.5 (a)], where confidence intervals for σ_r^2 are already given.

In order to give inference statements about the economic capital, simultaneous inference statements for the R risk categories are needed, e. g. simultaneous confidence statements of the form $\Pr(\boldsymbol{\rho} \in I_1 \times \ldots \times I_R) = 1 - \alpha$. In general, the construction of simultaneous confidence intervals is a challenging task, where approximations and inequalities such as the Bonferroni inequality are required. In the ASRF model, however, a single source of randomness influences all stochastic default probabilities in the same direction. This property is called comonotonicity, see [12, p. 199]. Consequently, it is possible to derive simultaneous confidence intervals directly from individual confidence intervals, since the simultaneous distribution of the R different variance estimators is degenerate.

Theorem 2 *Let the ASRF model given by (1) hold for $T \geq 1$ periods. Then*

$$\frac{\hat{\sigma}_1^2}{\sigma_1^2} = \frac{\hat{\sigma}_2^2}{\sigma_2^2} = \ldots = \frac{\hat{\sigma}_R^2}{\sigma_R^2}. \tag{5}$$

By using this theorem, simultaneous confidence intervals for the parameter vector $\boldsymbol{\rho}$ can be derived, which are stated here for the first time to the best of our knowledge.

Theorem 3 *Let the assumptions of Theorem 1 hold. Then each R-dimensional stochastic interval $I_1 \times \ldots \times I_R$, with I_r taken from Theorem 1, is a simultaneous $(1-\alpha)$-confidence interval for $\boldsymbol{\rho}$, i. e.*

$$\Pr(\boldsymbol{\rho} \in I_1 \times \ldots \times I_R) = 1 - \alpha \quad \textit{for all } (\boldsymbol{\pi}, \boldsymbol{\rho}).$$

4 Conclusion

In factor models, asset correlations determine the dependence structure of default events and have together with the portfolio weights great impact on the economic capital. Consequently, the quantification of the estimation errors of asset correlations by using confidence intervals and the implementation of a meaningful sensitivity analysis for asset correlations are important. For this purpose, asset correlations have been estimated from transformed default rates (probits of default rates). For this it is assumed that the systematic risk factors for the different periods are stochastically independent. By using the special structure of the ASRF model with only one systematic risk factor in each period, the individual confidence intervals for asset correlations given in Theorem 1 provide the basis for simultaneous confidence intervals for all asset correlations given in Theorem 3. The advantage of these confidence statements is that they are given in closed-form and are not the result of time-consuming Monte Carlo simulations. They are exact in the sense that they have the desired coverage probability even for a short time series of default rates, since they do not rely on asymptotics for $T \to \infty$.[4] Moreover, the approach based

[4] For an alternative estimation approach based on $T \to \infty$ see for example [15].

on probits of default rates can be extended to derive confidence intervals for default and survival time correlations.[5]

The length of the confidence intervals mirrors the parameter uncertainty and is for a fixed confidence level only caused by the variability inherent in the data. In this sense, confidence intervals can be understood as a statistical approach to sensitivity analysis and can be used to identify meaningful stress scenarios for correlations. The stress level involved may be systematically varied by the variation of the confidence level used, where higher confidence levels lead to wider confidence intervals. In contrast, stress tests based on historical simulation are limited by the range of the observed data so that extrapolation beyond the observations is not possible, e. g. stress scenarios in the extreme tails of distributions can not be considered.

Appendix

Proof of Theorem 1

From the stochastic independence of $Z_1,\ldots,Z_T$, it follows that the random variables $Y_{1r},\ldots,Y_{Tr}$ from (2) are stochastically independent and $\frac{T\hat{\sigma}_r^2}{\sigma_r^2} \sim \chi^2_{T-1}$. This implies

$$\begin{aligned}
1-\alpha &= \Pr\left(\chi^2_{T-1,\alpha/2} \leq \frac{T\hat{\sigma}_r^2}{\sigma_r^2} \leq \chi^2_{T-1,1-\alpha/2}\right) \\
&= \Pr\left(\frac{T\hat{\sigma}_r^2}{\chi^2_{T-1,1-\alpha/2}} \leq \sigma_r^2 \leq \frac{T\hat{\sigma}_r^2}{\chi^2_{T-1,\alpha/2}}\right) \\
&= \Pr\left(\frac{\hat{\rho}_r}{1-\hat{\rho}_r}\frac{T}{\chi^2_{T-1,1-\alpha/2}} \leq \frac{\rho_r}{1-\rho_r} \leq \frac{\hat{\rho}_r}{1-\hat{\rho}_r}\frac{T}{\chi^2_{T-1,\alpha/2}}\right) \\
&= \Pr(\rho_r \in I_r),
\end{aligned}$$

where the penultimate equality follows from (4) and the second equation in (3).

Proof of Theorem 2

Equations (2) and (3) imply $(Y_{t1}-\mu_1)/\sigma_1 = \ldots = (Y_{tR}-\mu_R)/\sigma_R = -Z_t$ for $t = 1,\ldots,T$. This implies $(\bar{Y}_1-\mu_1)/\sigma_1 = \ldots = (\bar{Y}_R-\mu_R)/\sigma_R$. Squaring and summation of the differences $(Y_{tr}-\bar{Y}_r)/\sigma_r = (Y_{tr}-\mu_r)/\sigma_r - (\bar{Y}_r-\mu_r)/\sigma_r$ yields $\sum_{t=1}^T (Y_{t1}-\bar{Y}_1)^2/\sigma_1^2 = \ldots = \sum_{t=1}^T (Y_{tR}-\bar{Y}_R)^2/\sigma_R^2$ and therefore (5).

Proof of Theorem 3

By using the proof of Theorem 1 together with Theorem 2, it follows that

[5] See [11].

$$\begin{aligned}\Pr(\boldsymbol{\rho} \in I_1 \times \ldots \times I_R) &= \Pr\left(\bigcap_{r=1}^{R}\left\{\chi^2_{T-1,\alpha/2} \leq \frac{T\hat{\sigma}_r^2}{\sigma_r^2} \leq \chi^2_{T-1,1-\alpha/2}\right\}\right)\\ &= \Pr\left(\chi^2_{T-1,\alpha/2} \leq \frac{T\hat{\sigma}_1^2}{\sigma_1^2} \leq \chi^2_{T-1,1-\alpha/2}\right)\\ &= 1-\alpha.\end{aligned}$$

References

1. Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards – A revised framework, comprehensive version. http://www.bis.org/publ/bcbs128.htm, June 2006. Accessed November 12, 2010.
2. Christian Bluhm, Ludger Overbeck, and Christoph Wagner. *Introduction to Credit Risk Modeling*. Chapman & Hall/CRC, Boca Raton, 2nd edition, 2010.
3. Jens H. E. Christensen, Ernst Hansen, and David Lando. Confidence sets for continuous-time rating transition probabilities. *J. Banking Finance*, 28(11): 2575–2602, 2004.
4. Klaus Düllmann and Monika Trapp. Systematic risk in recovery rates – an empirical analysis of U.S. corporate credit exposures. Number 2. Frankfurt am Main, June 2004.
5. R. Frey and A. J McNeil. Dependent defaults in models of portfolio credit risk. *Journal of Risk*, 6(1): 59–92, 2003.
6. Michael B. Gordy. A comparative anatomy of credit risk models. *J. Banking Finance*, 24(1–2): 119–149, 2000.
7. Michael B. Gordy. A risk-factor model foundation for ratings-based bank capital rules. *J. Finan. Intermediation*, 12(3): 199–232, 2003.
8. Samuel Hanson and Til Schuermann. Confidence intervals for probabilities of default. *J. Banking Finance*, 30(8): 2281–2301, 2006.
9. Steffi Höse. *Statistische Genauigkeit bei der simultanen Schätzung von Abhängigkeitsstrukturen und Ausfallwahrscheinlichkeiten in Kreditportfolios*. Shaker Verlag, Aachen, 2007.
10. Steffi Höse and Stefan Huschens. Simultaneous confidence intervals for default probabilities. In Martin Schader, Wolfgang Gaul, and Maurizio Vichi, editors, *Between Data Science and Applied Data Analysis*, pages 555–560. Springer, Berlin, 2003.
11. Steffi Höse and Stefan Huschens. Confidence intervals for correlations in the asymptotic single risk factor model. *Dresdner Beiträge zu Quantitativen Verfahren*, 50/09, 2009.
12. Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, 2005.
13. Alexander J. McNeil and Jonathan P. Wendin. Bayesian inference for generalized linear mixed models of portfolio credit risk. *J. Empirical Finance*, 14(2): 131–149, 2007.
14. Katja Pluto and Dirk Tasche. Estimating probabilities of default for low default portfolios. In Bernd Engelmann and Robert Rauhmeier, editors, *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing*, pages 79–103. Springer, Berlin, 2006.
15. Daniel Rösch and Harald Scheule. Stress-testing credit risk parameters: an application to retail loan portfolios. *The Journal of Risk Model Validation*, 1(1): 55–75, Spring 2007.
16. Philipp J. Schönbucher. *Credit Derivatives Pricing Models: Models, Pricing and Implementation*. John Wiley & Sons, Chichester, 2003.
17. Oldrich Vasicek. Loan portfolio value. *Risk*, 15(12): 160–162, December 2002.
18. Oldrich Alfons Vasicek. Limiting loan loss probability distribution. http://www.moodyskmv.com/research/files/wp/Limiting_Loan_Loss_Probability_Distribution.pdf, 1991. Accessed November 12, 2010.

Valuation of Complex Financial Instruments for Credit Risk Transfer

Alfred Hamerle and Andreas Igl

1 Introduction

The fair valuation of complex financial products for credit risk transfer (CRT) can provide a good basis for sustained growth of these markets and their recovery after the current financial crisis. Therefore, the risks of these structured credit securities (such as Collateralized Debt Obligations (CDO) and Credit Default Swap-Index tranches) have to be known as well as the investor's current risk aversion.

Many (even sophisticated) investors rely solely on agencies' ratings for the risk assessment and the valuation of CRT-products due to an information asymmetry between the originators and them. The use of an identical rating scale both for structured products like tranches and corporate securities like bonds tempted many investors to apply identical risk profiles to all products with identical ratings. However, the risk characteristics of CDO tranches differ significantly from comparably rated corporate bonds in relation to systematic risk. Additionally, investors assign different prices to equal cash-flows depending on their risk aversions. Due to the high marginal utility of cash-flows in bad economic times these should have higher weights in a risk valuation approach than income in a benign market environment.

In this article we focus our study on the quite liquid and transparent market of the CDS-Index "iTraxx Europe" and related tranches. We compare market spreads of the tranches with spreads obtained from (I) a simple valuation model integrating the systematic risk sensitivity of tranches and (II) an extended valuation model additionally integrating the investor's risk aversion. Based on our economical reasoning valuation models we obtain significantly differing prices for the investigated complex financial instruments for CRT compared to the market quotations.

Prof. Dr. Alfred Hamerle, Andreas Igl
Department of Statistics, Faculty of Business Management, Economics and Management Information Systems, University of Regensburg, 93040 Regensburg, Germany, e-mail: alfred.hamerle@wiwi.uni-regensburg.de, andreas.igl@wiwi.uni-regensburg.de. We thank Mr. Dr. Liebig and Mrs. Mrasek (*Deutsche Bundesbank*) as well as Mr. Gruber, Mr. Fuechsl and Mrs. Hu (*Bayerische Landesbank*) for providing the data sets within a research cooperation.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_19,

2 A Framework for Credit Modeling

Collateral pools of complex financial instruments for CRT are composed of simple assets such as loans, bonds or CDS contracts. Their default behaviour forms the basis for the risk characteristics of multi-name derivatives such as CDOs, CDS-Indices and STCDO. In this paper the default behaviour of the collateral pool's assets follows a Gaussian single risk factor model[1].

Rating-based risk measurement of the collateral pool and the tranches

CRT-products transform the credit risk of the underlying collateral pool into a set of securities with different risk profiles by using credit enhancements like subordination. These tranches represent discriminative parts of the asset pool's loss distribution. A specific tranche incurs losses only if the loss of the collateral pool exceeds the respective subordination level (attachment point). A loss realisation of the asset pool higher than the upper limit (detachment point) of the tranche leads to its total wipe-out.

The rating-based risk measurement of assets in the collateral pool as well as related structured products like tranches may depend on their unconditional expected loss. Using an identical rating scale for both, many investors were tempted to apply similar risk profiles to all products with identical rating grades. Extensive simulation studies show significantly different risk characteristics of CDO tranches in relation to systematic risk (e.g. a macroeconomic factor) compared to corporate bonds with identical unconditional expected loss and therefore equal rating[2]. Figure 1 compares the conditional expected loss "profiles" (conditional upon some systematic market factor M; $\mathbb{E}[L_{tr}|M]$) of a mezzanine tranche with a corporate bond, both with equal expected (loss) rating.

The analysis of conditional expected loss "profiles" (CEL curve) clearly points out that structured products (tranches) react much more sensitively to changes in the macroeconomic factor. Given a critical range of systematic risk factor realisations, $M \in [-6;3]$, the curve of the tranche rises much more steeply than for a corporate bond with comparable rating. The differing impact of systematic risk on financial products leads to consequences both in risk management and in risk valuation.

A "Bond Representation" Approach for Structured Instruments

In the "bond representation" approach we consider the structured instruments as single-name credit products such as bonds. Therefore, we fit the risk parameters of the "virtual" bond by using the single risk factor model in order to achieve a

[1] The derivation of this model from a Merton asset-value model is described e.g. in [6]. The basic asset-value model relates to the findings of Merton in [8].

[2] Extensive simulation studies were performed e.g. in [5] and [6].

preferably good approximation of the tranche's real default behaviour (CEL curve) as well as a good conformity of all risk characteristics.

For the approximation we assume a constant $LGD_{tr}^{(b)}$. The expected loss profile in the bond model is then given by

$$\mathbb{E}\left[L_{tr}^{(b)}|M\right] = p_{tr}^{(b)}(M) \cdot LGD_{tr}^{(b)} = \Phi\left(\frac{c_{tr}^{(b)} - \sqrt{\rho_{tr}^{(b)}} \cdot M}{\sqrt{1-\rho_{tr}^{(b)}}}\right) \cdot LGD_{tr}^{(b)}. \qquad (1)$$

For all non-senior tranches (with detachment point $<$ maximum loss in the collateral pool) we set $LGD_{tr}^{(b)} = 1$. We adapt the threshold $c_{tr}^{(b)}$ to ensure that the unconditional expected loss of the bond model equals the simulated unconditional expected loss of the tranche, $\mathbb{E}\left[L_{tr}^{(b)}\right] = \mathbb{E}[L_{tr}]$.

Furthermore, a nonlinear least squares method between the realized $\mathbb{E}[L_{tr}|M]$ and the approximated $\mathbb{E}\left[L_{tr}^{(b)}|M\right]$ is used as the fitting criteria of our "bond representation" approach. The estimated value for $\rho_{tr}^{(b)}$ equals the (implied) asset correlation of the tranche and measures its sensitivity concerning changes of the systematic risk factor.

3 Valuation

(I) A Simple Pricing Model Based on the CAPM and a Merton Asset-Value Model

Based on the results of the tranche systematic risk measurement, $\rho_{tr}^{(b)}$, (unconditional) risk-neutral default probabilities of a tranche used for valuation can also be derived from a Merton asset-value model. The (unconditional) risk-neutral default probability of a tranche is given by

$$q_{tr}^{(b)} = \Phi\left(c_{tr}^{(b)} + \sqrt{\rho_{tr}^{(b)}} \cdot \delta \cdot \sqrt{T}\right) \qquad (2)$$

where δ denotes the Sharpe ratio of the market[3] and T stands for the maturity of the financial instrument. Assuming a zero bond structure, the price of a bond (respectively an approximated tranche) with nominal $EAD = 1$ can be calculated as

$$B^d(0,T) = \exp^{-r \cdot T} \cdot \left(1 - \left(q_{tr}^{(b)} \cdot LGD_{tr}^{(b)}\right)\right). \qquad (3)$$

Here, r is the risk-free interest rate. Because of the market Sharpe ratio this valuation approach is based on the assumption of a constant at-the-money implied volatility.

[3] See [9] for an explanation of the Capital Asset Pricing Model (CAPM) and the Sharpe ratio of the market.

(II) A Pricing Model Using Option-Implied Risk Neutral Density Functions

In contrast, our second valuation approach also integrates the current risk aversion of investors. Based on the Arrow/Debreu theory[4], state prices observed on benign markets should be, due to their low marginal utility, far smaller than state prices observed on stressed markets. Therefore, we use the "DJ EURO STOXX 50" index as a proxy for the market factor M in order to derive state prices of different states of the economy.

On the basis of (daily) market index option prices the "volatility skew" is taken into account and a (daily) risk neutral density (RND) of M is deduced by means of the findings of Breeden and Litzenberger ([2]). The resulting implied risk-neutral density function is then included into the pricing of the CRT-products by

$$B^d(0,T) = \exp^{-r\cdot T} \cdot \int_{-\infty}^{\infty} \mathbb{E}\left[L_{tr}^{(b)}|M\right] f^{RND}(M)\, dM \tag{4}$$

where $f^{RND}(M)$ denotes the risk-neutral density function including the volatility skew (and therefore the risk aversion of the investors). In contrast to the normal density shape of an RND without volatility skew, our density has more probability-mass in the range of stressed states of the economy.

4 Empirical Results

The empirical results of this paper can be organized in two sections. *First*, we quantify the systematic risk of financial instruments for CRT by using our "bond representation" approach introduced in section 2. Table 1 shows key statistics of the estimated asset correlation $\rho^{(b)}$ for our time-series ("Series 7"), both for the CDS-Index iTraxx Europe (collateral pool) and the tranches[5].

It can be seen that the average values of $\bar{\rho}_{tr}^{(b)}$ are much higher for all tranches than those for comparable single-name instruments (e.g. $\bar{\rho}_{Pool}^{(b)} = 0.270009$, estimation using the "bond representation" approach for the collateral pool's assets) reflecting the dramatic increase of the tranche sensitivity to systematic risk[6]. Because of the price relevance of systematic risks the dramatically higher asset correlations of the tranches need to be compensated for by a significantly higher spread.

[4] See [1] and [4] for an introduction.

[5] In order to make market spreads and model spreads of the tranches comparable, we calibrate the daily default threshold $c_{Pool}^{(b)}$ of the collateral pool to approximately match the market spreads and the model spreads of the CDS-Index. For the (constant) asset correlation of the collateral pool we use a market standard value of $\rho_{Pool}^{(b)} = 27\%$, see e.g. [7].

[6] As described in [6], the thinner a tranche is the higher is its risk factor sensitivity. Alike the tranche asset correlations rise by increasing the number of assets in the collateral pool or by integrating assets with high systematic risk.

Table 1 Survey of the estimated asset correlations for the index and tranches by using the "bond representation" approach. Data: "On-the-run" period of the CDS-Index "iTraxx Europe series 7".

	Tranche width	Mean	Min	Max
Collateral pool	0% - 100%	0.270009	0.269629	0.270170
Tranche 1	0% - 3%	0.671976	0.666913	0.680902
Tranche 2	3% - 6%	0.899765	0.898934	0.901091
Tranche 3	6% - 9%	0.932790	0.931852	0.933808
Tranche 4	9% - 12%	0.946309	0.945451	0.947331
Tranche 5	12% - 22%	0.911069	0.910364	0.912060

Second, we compare both the simple valuation model assuming a constant volatility and the extended valuation model integrating the volatility skew as a proxy for the investor's risk aversion with the observed tranche spreads of the market. Figure 2 shows the time-series of the spreads for the CDS-Index as well as for the tranches 1, 2 and 5.

Comparing the tranche spreads of our extended valuation model (dark grey dashed dotted line) with the corresponding market spreads (black solid line) we obtain significantly higher model spreads for all non-equity tranches (tranches 2 to 5). Otherwise, the model spread for the equity tranche (tranche 1) is considerably lower than the market spread. Therefore, our findings are in line with [3].

Moreover, we observe model spreads from the simple valuation approach (light grey dashed line) which hardly differ from the market spreads for the senior tranche (tranche 5). On the other hand, model spreads for mezzanine tranches are higher than the observed market spreads, but lower than model spreads of the extended valuation approach (tranche 2)[7].

Our outcome indicates that the consideration of the higher systematic risk of tranches as well as the integration of the investors' current risk aversion was different in the market compared to our economical reasoning valuation models.

References

1. Kenneth Joseph Arrow. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies*, 31: 91–96, 1964.
2. Douglas T. Breeden and Robert H. Litzenberger. Prices of State-contingent Claims Implicit in Option Prices. *Journal of Business*, 51(4): 621–651, 1978.
3. Joshua D. Coval, Jakub W. Jurek, and Erik Stafford. Economic Catastrophe Bonds. *American Economic Review*, 99(3): 628–666, 2009.
4. Gérard Debreu. *Theory of value: An Axiomatic Analysis of Economic Equilibrium*. Cowles Foundation Monographs Series, 1959.
5. Martin Donhauser. *Risikoanalyse strukturierter Kreditprodukte: Dissertation Universität Regensburg*. 2010.

[7] As a result of our default threshold $c_{Pool}^{(b)}$ calibration, we find a quite good match of the market index spread and our two model index spreads (Index).

6. Alfred Hamerle, Thilo Liebig, and Hans-Jochen Schropp. Systematic Risk of CDOs and CDO Arbitrage. *Deutsche Bundesbank, Series 2: Banking and Financial Studies*, (13), 2009.
7. Christoph Kaserer and Tobias Berg. Estimating Equity Premia from CDS Spreads: EFA 2009 Bergen Meetings Paper, 2008.
8. Robert C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29(2): 449–470, 1974.
9. William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 19(3): 425–442, 1964.

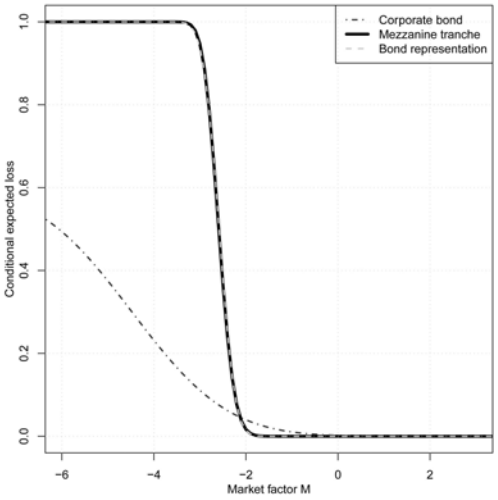


Fig. 1 Conditional expected loss profiles of a mezzanine tranche (black solid line) and a corporate bond (dark grey dashed dotted line) with equal rating. The figure also represents the goodness-of-fit of our "bond representation" approach (light grey dashed line) in contrast to a true CEL curve of a mezzanine tranche (black solid line).

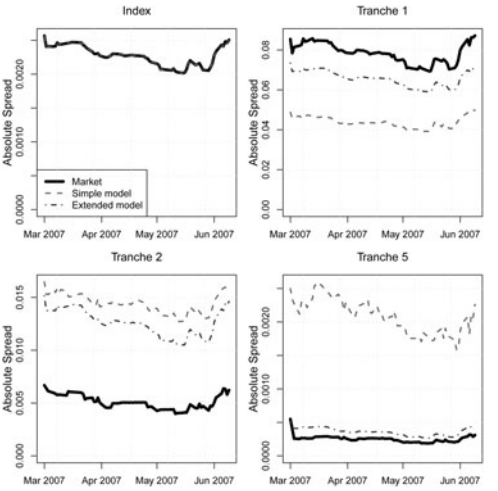


Fig. 2 Market spread (black solid line), model spread of the simple valuation model (light grey dashed line), and model spread of the extended valuation model (dark grey dashed dotted line) compared for the index as well as the tranches 1, 2 and 5.

VaR Prediction under Long Memory in Volatility

Harald Kinateder and Niklas Wagner

Abstract In their study on the applicability of volatility forecasting for risk management applications, [2] stress the importance of long-term volatility dependencies under longer forecast horizons. The present contribution addresses multiple-period value-at-risk (VaR) prediction for equity markets under long memory in return volatilities. We account for long memory in the τ-step ahead volatility forecast of GJR-GARCH(1,1) by using a novel estimator considering the slowly declining influence of past volatility shocks. Our empirical study of established equity markets covers daily index returns during the period 1975 to 2007. We study the out-of-sample accuracy of VaR predictions for five, ten, 20 and 60 trading days. As a benchmark model we use the parametric GARCH setting of Drost and Nijman (1993) and the Cornish-Fisher expansion as an approximation to innovation quantiles. The backtesting results document that our novel approach improves forecasts remarkably. This outperformance is only in part due to higher levels of risk forecasts. Even after controlling for the unconditional VaR levels of the competing approaches, the long memory GJR-GARCH(1,1) approach delivers results which are not dominated by the benchmark approach.

1 Introduction

Periods of financial market stress –including e.g. the market crash of October 1987, the burst of the internet bubble with accounting scandals at the beginning of the new millennium and the recent 2007/08 financial crisis– have increased both the

Harald Kinateder
Department of Business and Economics, Passau University, 94030 Passau, Germany, e-mail: Harald.Kinateder@uni-passau.de

Prof. Niklas Wagner
Department of Business and Economics, Passau University, 94030 Passau, Germany, e-mail: Niklas.Wagner@uni-passau.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_20,

regulatory as well as the industry demand for effective risk management. While financial institutions have to assure that their capital cushion is adequate *in advance of* a stress period such that institutions do not run into liquidity demands during a downturn, too conservative levels of risk capital lead to unnecessarily high levels of capital costs. In this setting, market volatility prediction plays a central role, where one of the "stylized facts" is the "long memory" (LM) property or the "long range dependence", which is typically observed by a slower than exponential decay in the autocorrelation function of, particularly absolute, asset returns; see [1], [3], [4] as well as [5], among others. As such, [2] stress the importance of long-term volatility dependencies under longer forecast horizons in their discussion of volatility forecasting for risk management applications.

In this contribution we use long memory based VaR approaches for multiple-period market risk prediction. First, we account for long memory in the τ-step ahead volatility forecast of the generalized autoregressive conditional heteroskedasticity (GARCH) model of [7], GJR-GARCH(1,1), by using a novel estimator considering the slowly declining influence of past volatility shocks. This is motivated by the fact that the influence of past volatility shocks on future volatility does not decline exponentially but rather hyperbolically. Second, based on the concept of self-affinity, we account for long memory by using a long memory based scaling approach for multiple-period volatility. As a benchmark model for multiple-period volatility, we use the parametric GARCH setting as introduced by [6]. As in [6], we model one-period returns on the, e.g., daily frequency and then use a prediction approach to derive multiple-period volatility rather than choosing the simpler approach of modeling aggregate multiple-period returns for volatility forecasting (via e.g. a standard GARCH model). This volatility prediction approach allows us to use daily return observations and increase forecasting ability.[1]

2 Multiple-Period Value-at-Risk

We assume that empirical returns[2] belong to a location-scale family of probability distributions of the form

$$R_t = \mu_t + \sigma_t Z_t,\ Z_t \sim i.i.d.\ F(0,1).$$

The location μ_t and the scale σ_t are $\mathcal{F}_{t-1}$-measurable parameters. In the setting above, with $\tau = 1,\ldots,h \in \mathbb{N}^+$, and $h < T$, the multiple-period return, $R_{t,t+h}$, is defined as $R_{t,t+h} \equiv \sum_{\tau=1}^{h} R_{t+\tau}$.

[1] See also the work of [9] who stress that, as far as parameter accuracy is concerned, a reduction in the number of observations appears not to be desirable.

[2] The one-period log return, R_t, is defined as $R_t \equiv \log P_t - \log P_{t-1}$. P_t denotes the price of a financial asset.

Assumptions:

1. F is a stationary distribution
2. $\text{Var}(R_{t,t+h}|\mathcal{F}_t) \equiv \sigma^2_{t,t+h} = \lambda\,\sigma^2_{t,t+1}$, with
   - $\lambda = h$ for *i.i.d.* processes and
   - $\lambda \in \mathbb{R}^+$ for LM processes

The VaR is defined as an upper threshold on losses in the sense that these losses are likely to exceed the value-at-risk with only a small probability α. $\mathcal{F}_t$ denotes the information set available up to date t. $F_{t,t+h}(\cdot)$ represents the conditional distribution of $R_{t,t+h}$ given information up to t, $F_{t,t+h}(r) = P(R_{t,t+h} \leq r|\mathcal{F}_t)$. Given some level α, the value-at-risk for $(t,t+h]$, $VaR^{\alpha}_{t,t+h}$, is defined as

$$VaR^{\alpha}_{t,t+h} = -\inf\left\{r : F_{t,t+h}(r) \geq \alpha\right\}, \tag{1}$$

with $0 < \alpha < 1$, and $h \geq 1$. This in turn yields

$$P(R_{t,t+h} \leq -VaR^{\alpha}_{t,t+h}|\mathcal{F}_t) = \alpha.$$

In our setting, the multiple-period VaR is

$$VaR^{\alpha}_{t,t+h} = -\big(\mu_{t,t+h} + \sigma_{t,t+h}F^{-1}(\alpha)\big), \tag{2}$$

where $\mu_{t,t+h}$ is the multiple-period mean and $\sigma_{t,t+h}$ is the multiple-period volatility of $R_{t,t+h}$. $F^{-1}(\alpha)$ is the quantile of the innovations distribution F.

3 Multiple-Period Volatility

3.1 Long Memory GJR-GARCH(1,1)

In the following we assume that one-period ahead returns are serially uncorrelated. It follows that

$$\sigma_{t,t+h} = \sum_{\tau=1}^{h} \sigma_{t+\tau}.$$

It is important to stress that the τ-step ahead volatility forecast for time $t+\tau$ should account for long range dependencies. Therefore, we use a novel estimator for τ-step ahead volatilities:[3]

$$\sigma_{t+\tau} = \sqrt{\sigma^2 + (\sigma_{t+1} - \sigma^2)g(\tau,H,\rho)}, \tag{3}$$

[3] The τ-step ahead GJR-GARCH(1,1) forecast is $\sigma_{t+\tau} = \sqrt{\sigma^2 + (\sigma_{t+1} - \sigma^2)(\alpha+\gamma+\beta)^{\tau-1}}$. The exponential decay of $(\alpha+\gamma+\beta)^{\tau-1}$ may not reflect the correct influence of past volatility shocks on future volatility.

where

$$g(\tau, H, \rho) = (\rho_{\tau-1})^{H-\rho_{\tau-1}}$$

under the following constraints:

$$1 \geq \rho_{\tau-1} \geq 0,$$

$$0 \leq g(\tau, H, \rho) \leq 1.$$

$\rho_{\tau-1}$ denotes the autocorrelation function of $|R_t|$, $H \in (0,1)$ is the Hurst exponent and σ^2 is the unconditional return variance. The function g ensures that past volatility shocks affect future volatility according to their magnitude of long-term dependence.

3.2 Long Memory Scaling

A common approach is to scale the one-period volatility by the square-root of time $\sqrt{h}$. This technique is only applicable when the underlying data is independent. Otherwise, the scaling formula would be misspecified. The degree of dependence can be calculated by the Hurst exponent, thus we use an unrestricted scaling exponent. As a result, the computation of volatility over h periods is as follows:

$$\sigma_{t,t+h} = h^H \sigma_{t,t+1}. \tag{4}$$

3.3 Drost/Nijman Formula

The multiple-period GARCH volatility forecast by [6] is:

$$\sigma_{t,t+h} = \sqrt{\omega_{(h)} + \alpha_{(h)} \varepsilon^2_{t-h,t} + \beta_{(h)} \sigma^2_{t-h,t}}, \tag{5}$$

where $\omega_{(h)}$, $\alpha_{(h)}$, and $\beta_{(h)}$ are the multiple-period parameters of a symmetric GARCH(1,1) process. Equation (5) requires an estimate for $\sigma^2_{t-h,t}$. If $\sigma^2_{t-h,t}$ is approximated badly, hence $\sigma^2_{t,t+h}$ is incorrect, as well. The Drost-Nijman technique does not use LM properties and may works badly when returns are dependent.

4 VaR Model Specifications

For $F^{-1}(\alpha)$ we use the skewed student-t distribution and the Cornish-Fisher expansion. The one-period ahead volatility, $\sigma_{t,t+1}$, is calculated by GJR-GARCH(1,1). In sum, we consider the following multiple-period VaR models:

Model 1: LM-based GJR-GARCH(1,1) & Student-*t*

$$VaR^{\alpha}_{t,t+h} = -\left(\mu_{t,t+h} + \sqrt{h\sigma^2 + (\sigma^2_{t+1} - \sigma^2)\sum_{\tau=1}^{h} g(\tau, H, \rho)}\; F^{-1}_{\nu(t)}(\alpha)\right)$$

Model 2: Drost/Nijman & Student-*t*

$$VaR^{\alpha}_{t,t+h} = -\left(\mu_{t,t+h} + \sqrt{\omega_{(h)} + \alpha_{(h)}\varepsilon^2_{t-h,t} + \beta_{(h)}\sigma^2_{t-h,t}}\; F^{-1}_{\nu(t)}(\alpha)\right)$$

Model 3: LM-Scaling & Student-*t*

$$VaR^{\alpha}_{t,t+h} = -\left(\mu_{t,t+h} + h^H \sigma_{t,t+1}\; F^{-1}_{\nu(t)}(\alpha)\right)$$

Model 4: Drost/Nijman & Cornish-Fisher Expansion

$$VaR^{\alpha}_{t,t+h} = -\left(\mu_{t,t+h} + \sqrt{\omega_{(h)} + \alpha_{(h)}\varepsilon^2_{t-h,t} + \beta_{(h)}\sigma^2_{t-h,t}}\; F^{-1}_{CF}(\alpha)\right)$$

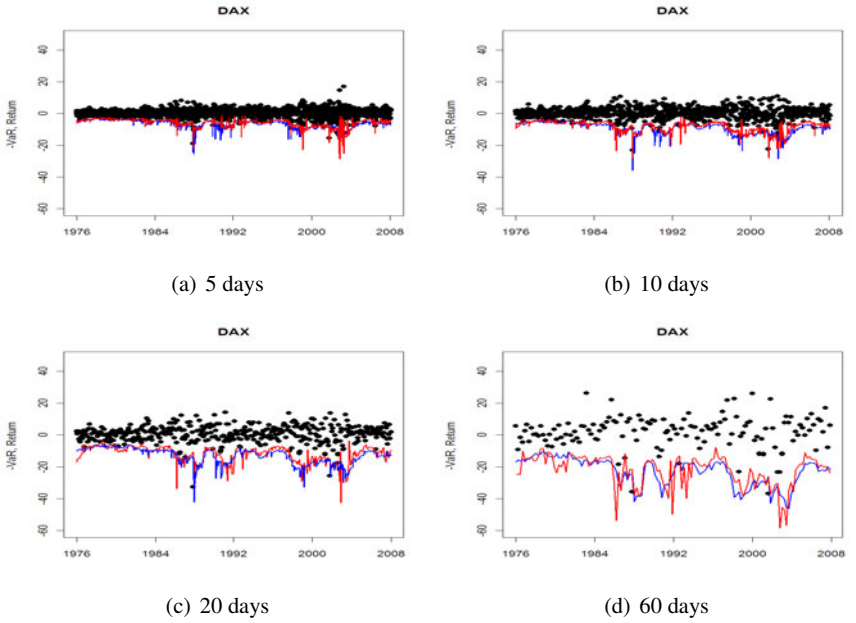
5 Empirical Analysis

The data is obtained from Thomson One Banker. It comprises 8,609 daily closing levels P_t from January 2, 1975 to December 31, 2007 of DAX, DOW JONES, NASDAQ, and S&P 500. In our study, we use non-overlapping continuously compounded percentage returns for different sampling frequencies $h \geq 1$.

Our results indicate that there is no significant advantage of using the Drost-Nijman formula for volatility computation. Model 1 achieves the best conditional coverage and lowest standard deviation of VaR. Figure 1 illustrates the VaR results for model 1 in comparison to our benchmark approach (model 4). For 60-day VaR forecasts the lower variability of VaR is clearly visible. For short horizons both LM models outperform the Drost/Nijman models substantially.

Additionally, we check how the models perform after calibration all approaches to the same average capital employed, that is to the same average sample VaR level of model 4.[4] By applying this criterion, we could approve that the different model performance is not only achieved due to different VaR levels, but also due to sophisticated adaption at volatility clusters. Even after recalibration model 1 outperforms the benchmark model significantly. More details can be found in [8]. To conclude, long memory in financial returns is a very important stylized fact which should be used for enhancing long-term VaR forecasts.

[4] Yet, we would like to stress that this test is an ex post examination, because the required capital is not known at the time of the estimate.



(a) 5 days (b) 10 days

(c) 20 days (d) 60 days

Fig. 1 Multiple-period VaR forecast for the DAX from January 2, 1976 to December 31, 2007. Model 1 is the blue line and model 4 is the red line.

References

1. T. G. Andersen and T. Bollerslev. Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance*, 52: 975–1005, 1997.
2. T. G. Andersen and T. Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4): 885–905, 1998.
3. R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1: 223–236, 2001.
4. Z. Ding, R. F. Engle, and C. W. J. Granger. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1: 83–106, 1993.
5. Z. Ding and C.W.J. Granger. Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics*, 73: 185–215, 1996.
6. F. C. Drost and T. E. Nijman. Temporal aggregation of GARCH processes. *Econometrica*, 61: 909–927, 1993.
7. Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48: 1779–1801, 1993.
8. H. Kinateder and N. Wagner. Market risk prediction under long memory: When VaR is higher than expected. Working paper, Passau University, 2010.
9. C.-M. Wong and K. P. So. On conditional moments of GARCH models, with applications to multiple period value at risk estimation. *Statistica Sinica*, 13: 1015–1044, 2003.

Solving an Option Game Problem with Finite Expiration: Optimizing Terms of Patent License Agreements

Kazutoshi Kumagai, Kei Takahashi, and Takahiro Ohno

Abstract In this study, we investigate the effect of expiration on the outputs of an option game problem. For this purpose, we solve a problem with finite expiration. Many previous studies have investigated the option game approach, and most of them consider infinite expiration in order to enable the problem to be solved analytically. However, there exist several situations in which it is necessary to consider finite expiration, e.g. a patent licensing agreement. Therefore, in this study, we focus on a model based on the case of a patent licensing agreement for a licenser and perform calculations for optimizing the agreement under the condition of finite expiration. We also determine the effect of sudden death risk on the outputs of an option game problem, which, in this case, are optimal terms for the licenser. By numerical experiments, we find that both expiration and sudden death risk have a strong influence on the optimal terms.

1 Introduction

As a result of previous studies by Dixt and Pindyck [1], there have been rapid developments in the real options approach. The real options approach developed in

Kazutoshi Kumagai
Department of Industrial & Management Systems Engineering, Graduate School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: k.kumagai@toki.waseda.jp

Kei Takahashi
Department of Industrial & Management Systems Engineering, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: k-takahashi@aoni.waseda.jp

Takahiro Ohno
Department of Industrial & Management Systems Engineering, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: ohno@waseda.jp

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 129
DOI 10.1007/978-3-642-20009-0_21,

consideration of the uncertainty in assessment of project value has attracted considerable attention as a valuation method in place of the conventional valuation parameter net present value (NPV).

However, the conventional traditional real options approach assumes that the decision making of a player is not affected by actions of other players. Then, many studies point out that the traditional real options approach overcounts the value of a project. To overcome this problem, the concept of game theory is incorporated into the real options approach; the resultant advanced method is termed option game approach.

Many previous studies have been conducted on the option game approach, e.g. Grenadier [2][1]. However, no study has yet solved a problem with finite expiration. Though expiration is an important factor for option pricing, many studies assume infinite expiration, because it enables to be solved the problem analytically. However, in many situations, it is necessary to consider the existence of expiration, such as a patent licensing game; in this paper we focus on this situation.

Patents usually have expiration, which has a strong influence on the terms of agreement. In this study, we model negotiations of a patent licensing agreement between a licenser and a licensee as an option game. After that, we optimize the terms of agreement for the licenser, show how to solve a problem with finite expiration, and demonstrate how expiration affects valuation. In following section, we discuss negotiations for patent licensing and optimization of contraction for the licenser by numerical methods.

2 The Situation

In this study, we consider negotiations between a risk-averse licenser, such as a university, and a risk-neutral licensee, such as an industrial firm. The negotiations are conducted for setting terms of patent licensing agreement.

A patent licensing agreement usually consists of two parts of license fees (R,F). One is a variable fee RS_t, called royalty, which is estimated by multiplying the royalty rate R by production that is based on the patent S_t at time t. The other part is a fixed fee F, called lump-sum payment, which is paid at the time of signing the contract. In this situation, we assume that once the agreement is accepted, the terms of agreement are valid until the patent expiration T.

The patent licensing agreement is influenced by the uncertainty of sales, initial investment for production I, cost percentage C, and other parameters. By modeling and optimizing this situation for the licenser, we calculate the optimal terms (R,F) for maximizing the licenser's expected utility V_P.

[1] Grenadier [2] modelled a pre-emption game in the real estate market as an option game problem. The game assumes an oligopolistic market consisting of two players, and the decision making of each player is affected by the other player's action.

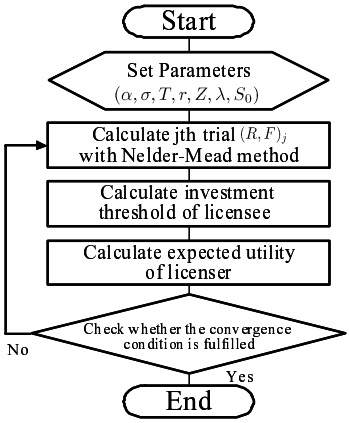


Fig. 1 Flowchart of general representation of calculation procedure adopted in this study.

3 The Model

The negotiation can be divided into the following two phases. and strategies which each player can take are as follows.

1. The licenser offers a pair of terms (R_t, F_t) at time t.
2. The licensee evaluates the expected value on the basis of (R_t, F_t) and incorporates its entry timing into the agreement T_0.

In the following, we solve the game based on a subgame perfect equilibrium and derive the optimal strategy of the licenser (R^*, F^*) by iterative calculation. Fig. 1 shows a general representation of the model considered in this study.

After this section, we present the calculation of the licenser's expected utility and consider each calculated utility as a trial. We assume that the sales S_t at time t is expressed by the following geometric Brownian motion:

$$\mathrm{d}S_t = \alpha S_t \mathrm{d}t + \sigma S_t \mathrm{d}z, \tag{1}$$

where α is the expected growth rate of S, σ is volatility, and $\mathrm{d}W$ is an increment of a standard Wiener process.

First, we represent the decision-making process of the licensee. Since the licensee is risk neutral, it acts to maximize the expected payoff V_L represented as

$$V_L \equiv \mathrm{E}\left[\int_{T_0}^{T} e^{-ru}\{(1-C-R)S_u - F\}\,\mathrm{d}u - Ie^{-rT_0}\right], \tag{2}$$

where r is the risk-free rate. By controlling the entry timing T_0, the licenser maximizes V_L. Second, we represent the licenser's action. We assume the licenser's utility function as follows:

$$U(x) = -\frac{1}{b}\exp(-bx), \tag{3}$$

where b is a parameter representing the degree of risk aversion. When the license agreement is valid, the licenser receives $RS_t + F - Z$ in each period, where Z is the cost of sustaining patents, which is paid to the patent office. Meanwhile, when it is invalid, the cost becomes $-Z$. Since the licenser is risk averse, it maximizes its expected utility V_P:

$$V_P \equiv \mathrm{E}\left[\int_0^{T_0} U(-Z)\,\mathrm{d}u + \int_{T_0}^{T} U(RS_u + F - Z)\,\mathrm{d}u\right]. \tag{4}$$

By controlling a pair of offering terms (R,F), the licenser acts to maximize V_P. As stated in section 1, there exists a particular risk in patent licensing, called sudden death risk. By modelling sudden death risk as a Poisson process, with a Poisson intensity of λ, we can manage the risk by adding λ to the discount rate. [2]

Using Ito's lemma, we can derive the following partial differential equations (PDEs) from (2) and (4).

$$(1-C-R)S_t - F - rV_L + \frac{\partial V_L}{\partial t} + \alpha S_t \frac{\partial V_L}{\partial S_t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V_L}{\partial S_t^2} = 0 \tag{5}$$

$$-\frac{1}{b}\exp(b(RS_t + F - Z_t)) - rV_P + \frac{\partial V_P}{\partial t} + \alpha S_t \frac{\partial V_P}{\partial S_t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V_P}{\partial S_t^2} = 0 \tag{6}$$

To derive V_L, V_P and optimize terms of agreement (R,F), we employ a combination of the Nelder-Mead method and the finite differential method (FDM) [3] Nelder-Mead method employs a figure which called a simplex, which is a figure having $K+1$ corners optimization is done in K dimensions. The Nelder-Mead method has an advantage over other optimization methods such as the Newton method. The Nelder-Mead method can optimize an objective function without deriving it.[4] FDM is a numerical calculation method usually used for option valuation in financial field. We solve (5) using the Crank-Nicolson scheme under boundary conditions commonly used for calculating the investment threshold of the licensee at time t: S_t^*. After that, using S_t^*, we solve (6) by the explicit scheme. We employ this scheme because the licencer's utility is discontinuous with S, and cannot be solved by the Crank-Nicolson scheme. In addition, while the time step is set to Δt for solving (5), it must be set to Δt^2 for solving (6), since the calculation errors for the explicit and Crank-Nicolson schemes are different.

[2] Schwartz [5] states that we need to consider sudden death risk when modelling patents licensing by the real options approach.

[3] According to the technique for approximating partial differentials, FDM is classified into three types - explicit scheme, implicit scheme, and Crank-Nicholson scheme. For more information about FDM, see Wilmott [6].

[4] For more information about the Nelder-Mead method, see Nelder and Mead [3].

4 Numerical Experiments

In this section, we discuss some numerical experiments conducted on the model. The parameters used in the experiments are explained in the figure captions.

Fig. 2 shows the effect of expiration T and sudden death risk intensity λ on the optimal terms of agreement (R^*, F^*). When the expiration T is longer, a variable fee is preferred to a fixed fee.

When the risk of the contract is significant, the licenser's investment threshold increases. Thus, the licenser has to offer more favorable terms that are acceptable for the licensee. This is because the longer the patent expiration, the higher is the risk of the contract.[5] Generically, it is said that a fixed fee is profitable for the licenser because it can shift the risk to the licensee.

Meanwhile, we acknowledge that sudden death risk also has a strong influence on the optimal terms, particularly on the fixed fee. Since fixed fee has already been paid to the licenser at the time of making the contract even when a sudden death risk event comes into being, the licenser prefers a decrease in the fixed fee to that in the variable fee.

Fig. 1, shown on the next page, shows the percentage of fixed and variable fees in the total fee paid during the making of the contract. When expiration T is short, the risk of sales fluctuating significantly is low. Therefore, a large fixed fee is favorable and preferable for the licenser.

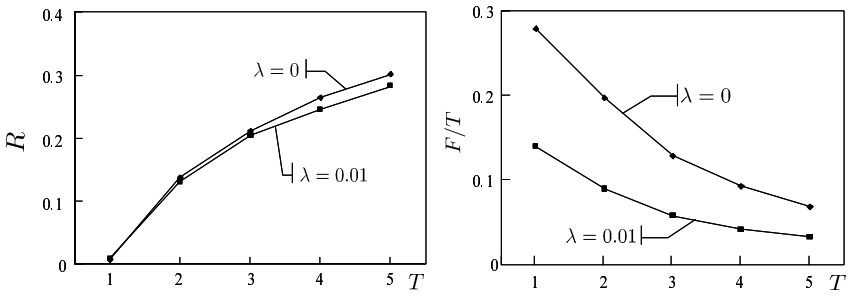


Fig. 2 Effect of expiration T and sudden death risk intensity λ on optimal terms of agreement (R^*, F^*). $\lambda = 0.01$ means a sudden death risk event come into existence 1 percent a year. $1 - \exp(-\lambda * t) = 1 - \exp(-0.01 * 1) = 0.01$ In this figure, we use the following parameters. $r = 0.03$, $\alpha = 0.03$, $\sigma = 0.1$, $S_0 = 10$, $I = 3$, $C = 0.5$, $b = 5.0$, and $Z = 1$.

[5] Saracho [4] modelled a patent licensing situation using game theory; he stated that when industrial competition is intense, a variable fee is more preferable for the licenser under optimal terms of agreement.

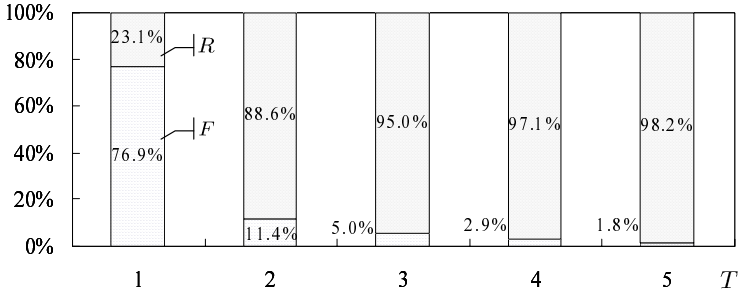


Fig. 3 Percentage of fixed and variable fees in total fee paid during making of contract. The parameters used in this figure have the same values as those mentioned in the caption of Fig. 2; $\lambda = 0$.

5 Conclusion and Future Tasks

In this study, we solved an option game problem by considering the case of finite expiration; this is because there exist several situations in which the consideration of expiration is essential. The focus of this study was on patent licensing agreements. We conducted numerical experiments and found that expiration has a strong influence on the outputs of option game problems.

We also showed how the parameters of patent licensing affect the optimal terms of agreement for the licenser. When the expiration is longer, the percentage of variable fee increases drastically. Additionally, we showed that sudden death risk has a strong influence on the optimal terms. Even though the occurrence of a sudden death risk event is very low, it leads to degradation of the terms for the licenser.

In the future, certain tasks need to be performed to obtain better results for the model. For example, we should model the situation more realistically by accurately reproducing characteristics of a patent licensing agreement. Further, other additional terms apart from (R, F) considered in this study need to be addressed as well.

References

1. A. K. Dixit and R. S. Pindyck. *Investment Under Uncertainty*. Princeton Univ Pr, 1994.
2. S. R. Grenadier. The Strategic Exercise of Options: Development Cascades and Overbuilding in Real Estate Markets. *The Journal of Finance*, 51(5): 1653–1679, December 1996.
3. J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4): 308–313, 1965.
4. A. I. Saracho. Patent Licensing Under Strategic Delegation. *Journal of Economics & Management Strategy*, 11(2): 225–251, 2002.
5. E. S. Schwartz. Patents and R&D as Real Options. *Economic Notes by Banca Monte dei Paschi di Siena SpA*, 33(1): 23–54, 2004.
6. P. P. Wilmott. *Paul Wilmott Introduces Quantitative Finance*. Wiley, U.S.A., 2007.

I.5 Pricing and Revenue Management

Chair: Prof. Dr. Alf Kimms (Universität Duisburg-Essen)

Revenue Management covers models and methods for accepting or rejecting demand especially in cases where demand is non-deterministic and capacity is not fully flexible in the short run. Pricing is closely connected, because customers have a limited willingness-to-pay. Applications of this kind occur in many industries, service as well as manufacturing. Revenue Management brings together classical fields like production, logistics, marketing and IT management. Quantitative models and methods for forecasting, capacity control, dynamic pricing, and overbooking are needed to attack such problems. Aspects like customer choice behaviour and competitors with their own revenue management systems, for example, make real situations become very complex. Specific applications often require tailor-made solutions.

We seek for theoretical papers as well as case studies where Operations Research techniques are used to contribute to the problem.

Switching Times from Season to Single Tickets

Ertan Yakıcı and Serhan Duran

Abstract In this paper, we developed the basic problem with only one switch from selling bundles to selling single tickets to a new version where "early switch to the low-demand event" is also allowed. The resulting policy is defined by a set of threshold pairs of times and remaining inventory, which determine the timing of the optimal switches.

1 Introduction

In sports and entertainment (S&E) industry, one of the most important market segmentations is that some customers buy season packages, or bundles of tickets to events during the season, while others buy individual tickets. Season ticket holders are important to the success of the organization, since they are more likely to donate to the organization or renew tickets in the future.

As the common S&E industry practice, first season tickets are put on the sale, and the individual tickets follow them by a time lag. In some cases, this switching date is announced in advance [1], but in many other cases the date is announced after the start of the selling season. For example, the Atlanta Hawks basketball team announced by email when single tickets went on sale for the 2008-2009 season. A key reason for this practice is simple: the organization can adapt the switching time to the realization of demand and improve overall profit. The time uncertainty in starting the sales of single tickets by the organization can also encourage customers to buy season tickets to ensure attending popular events or obtaining good seats.

Ertan Yakıcı
Middle East Technical University, e-mail: e156477@metu.edu.tr

Serhan Duran
Middle East Technical University, e-mail: sduran@ie.metu.edu.tr

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_22,

The model developed by [2] serves as a reference to the study reported in this paper. They studied the specific question of timing the single switch from selling bundles to selling single tickets. In this paper, we developed the basic problem to a new version where "early switch to the low-demand event" is also allowed. The resulting policy is defined by a set of threshold pairs of times and remaining inventory, which determine the timing of the optimal switches. After each sale, if the current time is less than the corresponding time threshold, then the switch is exercised.

2 The Dynamic Timing Problem

2.1 Assumptions

Let $M \in \mathbb{Z}^+$ be the number of seats available for sale at the venue where the events will take place, and $T \in \mathbb{R}^+$ be the selling period. There are two events that will be performed at the venue; a high-demand and a low-demand event. We focus on the selling horizon before the season begins and assume that the selling period ends when the first event takes place. The selling period begins with first offering tickets as a bundle (one high-demand and one low-demand event together) at price p_B, then switching to selling low-demand event tickets at p_L, and the bundle at p_B and then finally switching to selling only individual event tickets at p_L and p_H for individual low-demand and high-demand events, respectively. We assume that the product prices are predetermined at the beginning of the selling season, which is true for most organizations, especially during the time preceding the start of the season.

We assume that for each product, there is a corresponding Poisson process of demand: $N_B(s)$, $0 \le s \le t$, with known constant intensity λ_B for the bundled events; $N_L(s)$, $0 \le s \le t$, with known constant intensity λ_L, and $N_H(s)$, $0 \le s \le t$, with known constant intensity λ_H for the low-demand and high-demand events, respectively. In addition to these three demand processes, we have the Poisson process of $N_{B'}(s)$, $0 \le s \le t$, with known constant intensity $\lambda_{B'} (< \lambda_B)$ for the demand of the *moderated* bundled tickets while low-demand tickets are on sale. The sharing of capacity among bundled tickets and low-demand event tickets when they are sold simultaneously is reflected by using a combined Poisson process, $(t, N_{B'L}(t))$, where $\{N_{B'L}(t),\, t \ge 0\} = \{N_{B'}(t) + N_L(t),\, t \ge 0\}$.

The state of the system is indicated by the elapsed time t and the remaining number of seats at time t for low-demand and high-demand events, $(n_L(t), n_H(t))$ or simply (n_L, n_H) if it is observable at time t. We define $r_L = \lambda_L p_L$ and $r_H = \lambda_H p_H$ as the revenue rate for low-demand and high-demand event ticket sales, respectively. The revenue rate for the bundled tickets when they are the only product on sale is $r_B = \lambda_B p_B$, and it reduces to $r_{B'} = \lambda_{B'} p_B$ when it is sold along with the low-demand event tickets. We assume that the expected revenue rates satisfy the relation: $r_B > r_{B'} + r_L > r_H + r_L$. Otherwise, switching immediately would be optimal for all states.

3 Model and Results

The expected total revenue over the time horizon $[t,T]$ with the optimal switching times τ_1 and τ_2 is given by $V_1(t,(n_L,n_H))$:

$$\begin{aligned} V_1(t,(n_L,n_H)) &= \sup_{\tau_1\in\mathcal{T}} E\big[p_B\big((N_B(\tau_1)-N_B(t))\wedge n_L\big)\big] \\ &\quad + V_2(\tau_1,(n_L(\tau_1),n_H(\tau_1))), \textit{ where} \qquad (1) \\ V_2(t,(n_L,n_H)) &= \sup_{\tau_2\in\mathcal{T}} E\Big[p_L\big((N_{B'L}(\tau_2)-N_{B'L}(t))\wedge n_L\big)\frac{\lambda_L}{\lambda_{B'}+\lambda_L} \\ &\quad + p_B\big((N_{B'L}(\tau_2)-N_{B'L}(t))\wedge n_L\big)\frac{\lambda_{B'}}{\lambda_{B'}+\lambda_L}\Big] \\ &\quad + V_3(\tau_2,(n_L(\tau_2),n_H(\tau_2))), \textit{ and} \qquad (2) \\ V_3(t,(n_L,n_H)) &= E\Big[p_L\big((N_L(T)-N_L(t))\wedge n_L\big)+p_H\big((N_H(T)-N_H(t))\wedge n_H\big)\Big] \qquad (3) \end{aligned}$$

where $\mathcal{T}$ is the set of switching times τ_k satisfying $t\le\tau_1\le\tau_2\le T$ and $(x\wedge y)$ indicates the minimum of the x and y. The remaining number of seats for sale at the first switch (just before starting selling low-demand event tickets) are given by $n_L(\tau_1)=[n_L-N_B(\tau_1)+N_B(t)]^+$ and $n_H(\tau_1)=[n_H-N_B(\tau_1)+N_B(t)]^+$, where $x^+=\max\{0,x\}$. The remaining number of seats for sale at the second switch (just after stopping selling bundles) are given by $n_L(\tau_2)=[n_L(\tau_1)-N_{B'L}(\tau_2)+N_{B'L}(t)]^+$ and $n_H(\tau_2)=[n_H(\tau_1)-((N_{B'L}(\tau_2)-N_{B'L}(t))\wedge n_L(\tau_1))\frac{\lambda_{B'}}{\lambda_{B'}+\lambda_L}]^+$. The expected total revenue is the sum of revenues from the bundled tickets sale before the first switch over $[0,\tau_1]$, from the bundle and low-demand event sale after the first switch until the second switch with $n_L(\tau_1)$ and $n_H(\tau_1)$ seats available for sale over $[\tau_1,\tau_2]$ for the low-demand and high-demand events, respectively, and the expected revenue from single tickets after the second switch with $n_L(\tau_2)$ and $n_H(\tau_2)$ items available for sale over $[\tau_2,T]$.

At time t, if we can compare the expected revenue over $[t,T]$ from switching immediately, to the expected revenue if we delay the switch to a later time τ_k ($t\le\tau_k\le T$), then we can decide whether delaying the switch further is beneficial or not. In order to do that we define the infinitesimal generator $\mathcal{G}_2$ with respect to the Poisson process $(t,N_{B'L}(t))$ for a uniformly bounded function $g(t,(n_L,n_H))$ when $n_L\ge 1$ and $n_H\ge 1$ as,

$$\begin{aligned} \mathcal{G}_2 g(t,(n_L,n_H)) &= \frac{\partial g(t,(n_L,n_H))}{\partial t}+\lambda_{B'}[g(t,(n_L-1,n_H-1))-g(t,(n_L,n_H))] \\ &\quad + \lambda_L[g(t,(n_L-1,n_H))-g(t,(n_L,n_H))], \end{aligned}$$

and define a similar infinitesimal generator $\mathcal{G}_1$ with respect to the Poisson process $(t,N_B(t))$ as,

$$\mathcal{G}_1 g(t,(n_L,n_H)) = \frac{\partial g(t,(n_L,n_H))}{\partial t}+\lambda_B[g(t,(n_L-1,n_H-1))-g(t,(n_L,n_H))].$$

Applying $\mathcal{G}_1$ and $\mathcal{G}_2$ to the functions $V_2(t,(n_L,n_H))$ and $V_3(t,(n_L,n_H))$ gives the immediate loss of revenue from low-demand event and moderated bundle sales together and only single tickets sales, respectively, if the corresponding switches are delayed. But during the time when switching is delayed, the current Poisson processes are active. Therefore, the net marginal gain (or loss) for delaying the second switch at state $(t,(n_L,n_H))$ is given by $\mathcal{G}_2V_3(t,(n_L,n_H))+\lambda_L p_L+\lambda_{B'}p_B$. Similarly, the net marginal gain (or loss) for delaying the first switch to selling bundles and low-demand games simultaneously at state $(t,(n_L,n_H))$ is given by $\mathcal{G}_1V_2(t,(n_L,n_H))+\lambda_B p_B$.

By Dynkin's Lemma [5] and the optional sampling theorem [4], we get that $V_k(t,(n_L,n_H))=V_{(k+1)}(t,(n_L,n_H))+\widetilde{V_k}(t,(n_L,n_H))$. We prove that this equation can be interpreted as the optimal revenue over $[t,T]$ consisting of two parts: the revenue after the k^{th} switch ($k=1,2$) and the additional revenue from delaying the k^{th} switch further in time. Moreover, when $\widetilde{V_k}(t,(n_L,n_H))=0$, delaying the switch further is not optimal, whereas $\widetilde{V_k}(t,(n_L,n_H))>0$ implies a revenue potential from delaying the switch.

To compute $\widetilde{V_1}((t,(n_L,n_H))$ and $\widetilde{V_2}((t,(n_L,n_H))$, we introduce the mirror functions $\overline{V_1}(t,(n_L,n_H))$ and $\overline{V_2}(t,(n_L,n_H))$, which can be derived recursively and are identical to $\widetilde{V_1}(t,(n_L,n_H))$ and $\widetilde{V_2}(t,(n_L,n_H))$, respectively. We have demonstrated the existence of the alternate function $\overline{V_1}(t,(n_L,n_H))$ and $\overline{V_2}(t,(n_L,n_H))$ similar to Theorem 1 in [3]. We will first focus on the second switch decision. The properties for it implies a similar policy for the first switch decision.

Theorem 1 *For all $1\le n_L\le n_H\le M$, and when $\lambda_{B'}>\lambda_H$, $\overline{V_2}(t,(n_L,n_H))$ is recursively determined by*

$$\overline{V_2}(t,(n_L,n_H))=\begin{cases}\displaystyle\int_t^T L_2(s,(n_L,n_H))e^{-(\lambda_L+\lambda_{B'})(s-t)}ds & \text{if } t>x^2_{(n_L,n_H)}\\ 0 & \text{otherwise,}\end{cases}\tag{4}$$

where

$$x^2_{(n_L,n_H)}=\inf\Big\{0\le t\le T:\ \int_t^T L_2(s,(n_L,n_H))e^{-(\lambda_L+\lambda_{B'})(s-t)}ds>0\Big\},$$

$$\begin{aligned}L_2(t,(n_L,n_H))&=\lambda_{B'}\overline{V_2}(t,(n_L-1,n_H-1))+\lambda_L\overline{V_2}(t,(n_L-1,n_H))\\&\quad+\mathcal{G}_2V_3(t,(n_L,n_H))+\lambda_L p_L+\lambda_{B'}p_B,\qquad 0\le t\le T,\end{aligned}$$

The proof for the Theorem is along with the lines of the proof of Theorem 2 in [3]. What we have shown so far is that for any inventory level (n_L,n_H), there exists a time $x^2_{(n_L,n_H)}$ such that: $\overline{V_2}(t,(n_L,n_H))>0$ if $t>x^2_{(n_L,n_H)}$, and $\overline{V_2}(t,(n_L,n_H))=0$ if $t\le x^2_{(n_L,n_H)}$. Therefore, if the system reaches remaining inventory level (n_L,n_H) at a time $t\le x^2_{(n_L,n_H)}$, then it is optimal for the second switch decision to be implemented immediately. On the other hand if it takes the system longer than $x^2_{(n_L,n_H)}$ time units to reach remaining inventory level (n_L,n_H), then it is optimal to delay the second switch further. Therefore, the $x^2_{(n_L,n_H)}$ values can be interpreted as the

latest switching time or the switching-time thresholds for the second switch, when (n_L, n_H) items are unsold.

$\overline{V_1}(t,(n_L,n_H))$ is recursively determined in a similar way to the mentioned in Theorem 1, and it is along with the lines of Theorem 3 in [2]. For any inventory level (n_L,n_H), there exists a time $x^1_{(n_L,n_H)}$ such that: $\overline{V_1}(t,(n_L,n_H)) > 0$ if $t > x^1_{(n_L,n_H)}$, and $\overline{V_1}(t,(n_L,n_H)) = 0$ if $t \leq x^1_{(n_L,n_H)}$. Therefore, if the system reaches remaining inventory level (n_L,n_H) at a time $t \leq x^1_{(n_L,n_H)}$, then it is optimal for the first switch decision to be implemented immediately. Moreover, we can easily see that the switching-time thresholds, $x^1_{(n_L,n_H)}$ and $x^2_{(n_L,n_H)}$ are non-increasing in unsold inventory n_L and n_H. Intuitively, as the unsold inventory increases, it is beneficial for the team to delay the switch further in order to take advantage of the bundle sales longer.

4 Computational Experiments

We consider a team with 120-ticket stadium facing the problem of selling tickets to one high-demand and one low-demand game during a selling season that lasts 2 months. The demand rates for the games are 50 and 40 seats per month and the prices to be charged for these seats are \$200 and \$50 for the high and the low-demand game, respectively. If the seats are sold as a bundle with one high and one low-demand seat (without any single game ticket sale at the same time), the demand rate will be 130 seats per month for the bundle. On the other hand, if the bundles are sold along with low-demand game ticket at the same time, the rate is reduced to 80 seats per month for the bundle. The price for the bundle ticket is assumed to be \$220.

The numerical experimentation demonstrates the impact on revenue of deciding the switch time dynamically instead of using a static switch time. The bundle and moderated bundle rates are decreased and increased by 10 arrivals per month in order to see the effects of demand rate changes on the revenue improvement. We create 100,000 random sample paths for the arrival of customers, and calculate the average revenue when the switch time is decided dynamically and statically for those paths. While comparing the dynamic decision policy with the static one, we divided the selling horizon of two months into 500 equal intervals and all possible combinations of static switch times are tried. We did not encounter any improvement less than 0.44%. Here, we present just a few selected combinations for brevity. Figure 1 illustrates the percentage revenue improvement of dynamic switching over the static case. The revenue improvement of the dynamic switching method over the static case for the selected static switch time combinations can be between 0.5-6.71% when the model parameters set as mentioned.

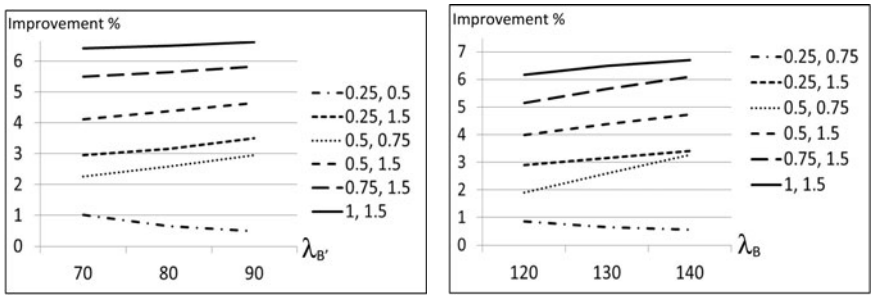


Fig. 1 Percentage Revenue Improvement of Dynamic Switching Over Static Case for Various First and Second Switch Time Pairs

5 Conclusions

In this paper, we have studied an extension of the basic problem of switching from selling bundles of tickets first to selling individual tickets afterwards so as to maximize revenue over a selling season [2]. The contribution of this study is its allowing the option of early switch to low-demand game tickets. We see that the option of early switch to low-demand game ticket can create up to 6.71% improvement.

References

1. M. Drake, S. Duran, P. Griffin, and J. Swann. Optimal timing of switches between product sales for sports and entertainment tickets. *Naval Research Logistics*, 55(1): 59–75, 2008.
2. S. Duran and J. Swann. Dynamic switching times from season to single tickets in sports and entertainment, Working Paper, Georgia Institute of Technology, 2009.
3. Y. Feng and B. Xiao. Maximizing revenues of perishable assets with a risk factor. *Operations Research*, 47(2): 337–341, 1999.
4. I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus.* Springer-Verlag, NY, 1988.
5. L.C.G.Rogers and D. Williams. *Diffusions, Markov Processes and Martingales, Volume 2: Itô Calculus*, John Wiley & Sons, NY, 1987.

A Revenue Management Slot Allocation Model with Prioritization for the Liner Shipping Industry

Sebastian Zurheide and Kathrin Fischer

Abstract The main characteristics for the successful use of revenue management are present in the liner shipping industry. Hence, revenue management methods can be applied to obtain the best container acceptance strategy. The segmentation which is normally used is based on the different container types, routes and customers. The literature shows that reliability and delivery speed are important factors for customers, and that containers with different commodities often have different priority. Therefore, the optimization model which is developed in this work creates booking limits for standard segmentations and incorporates an additional segmentation approach. These new segments are express and standard containers: The express segment is intended for urgent cargo and gets priority on the next ship. A container in the standard segment only needs to be shipped until a fixed delivery date and can be postponed. This new segmentation approach generates advantages for both parties involved.

1 Introduction

Liner shipping companies have to deal with the problem of assigning container slots to the right customer at the right price to maximize their revenue. In this situation, revenue management (RM) systems can help to find the revenue maximizing slot allocation strategy for each ship. Liner shipping companies offer different services connecting a certain number of ports. Here, the focus is on deep sea shipping services, where the services connect continents and most of the services offered are loops. A deep sea loop consists of a number of port calls at one continent followed by a number of port calls at another. [6, p. 50] These services operate with a fixed

Sebastian Zurheide and Kathrin Fischer
Institute for Operations Research and Information Systems, Hamburg University of Technology, Schwarzenbergstr. 95 D, 21073 Hamburg,
e-mail: zurheide@tu-harburg.de and kathrin.fischer@tu-harburg.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_23,

schedule, and in most cases a weekly departure is offered at each port. The busiest legs in these loops are the intercontinental connections. Especially on these sections of the route, capacity is tight and needs to be managed.

For the successful use of an RM system, [1, p. 39] identifies seven common characteristics: perishability, fixed capacity, low marginal and high fixed cost, segmentable demand structures, advance sales/bookings, stochastic and fluctuating demand, and knowledge of historic sales data and demand forecasting. These main characteristics for the use of RM can be found in the liner shipping industry: Because of the fixed departure times, container slots are perishable. The ships also have a fixed maximum slot capacity and a maximum deadweight. As a third characteristic, there are high fixed costs for operating the ship and only low costs for transporting an additional container. The demand can be separated by different container types, routes and into a spot market and a contractual market. The different container bookings have to be made in advance of the departure and the booking process starts weeks before the departure. The sixth main characteristic of a stochastic and fluctuating demand can also be found in the liner shipping business, as many service routes have to deal with variable demand, because of seasonal and/or economic variations. Finally, historical sales data are available or can be collected in modern computer based booking systems. With these data, forecasting is possible for all the different routes, types and segments.

Hence, on routes where demand exceeds the capacity, the relevant characteristics for RM are present. In the following, the well-established differentiations are described and an additional service-related segmentation is suggested. This segmentation then is incorporated into a quantitative slot allocation model.

2 Market Segmentation and Product Differentiation

The market segmentation and product differentiation as one of the seven characteristics mentioned above is an important part of an RM system. Currently, there are three main segmentation criteria used in the liner shipping market, which divide it by the different container types, routes and customers. The first and second criteria are physical conditions which create different products, and the third criterion is a customer condition to differentiate the customers.

2.1 The Well-Established Segmentation Criteria

The first segmentation by container type is based on the different standardized dimensions of containers defined in ISO 668. The most common types are 20’ Dry or Reefer, 40’ Dry or Reefer and 40’ Dry High Cube containers. The second differentiation criterion are the routes connecting different ports within the services. Each route has a different demand structure, and demand varies by season and by

direction. This also causes a major trade imbalance between many regions. Therefore, empty containers need to be relocated to assure equipment availability in those regions with more outbound than inbound traffic. The third differentiation criterion is based on the customers. Liner shipping companies have mainly two different types of customers, contractual and ad-hoc customers. Contractual customers sign agreements for a number of containers they are going to ship with the carrier in a specific time range. These contracts last for several months and guarantee the carrier a minimum quantity of containers. Spot market containers are those containers from ad-hoc customers who do not have a contract with the carrier. These customers do not have a fixed contract price or reserved slots and in most cases they are willing to pay a higher price. The different products and customers usually compete for the same capacity on the ship.

2.2 Service-Related Segmentation Approach

In addition to the product and customer segmentations described in 2.1, [5] suggest a segmentation based on the urgency of the shipment. Customers with urgent cargo are willing to pay more for its shipment, and it is important for these customers to get slots on a certain ship. In contrast to that, a customer with non-urgent cargo can wait for the next ship and try to find the cheapest freight rate. This segmentation approach is similar to the air cargo industry, where a "time definite" segmentation was introduced in 1998 by Lufthansa Cargo and adopted by other air cargo carriers. Air cargo customers at Lufthansa Cargo can choose from one standard and three express segments with different delivery time frames. [1, p. 11] Concerning the urgency of different shipments, [6, p. 520f] points out that different liner shipping customers have different priorities, mainly depending on their cargo. Especially for shippers with high value commodities, delivery speed and reliability are important factors. Also in surveys, these factors were identified as crucial for shippers in the liner shipping industry. [3]

Therefore, these ideas are used to create a slot allocation model with a time-based segmentation consisting of express and standard containers. The standard segment is designed for non-urgent cargo and the containers can be postponed if necessary. These postponed containers are assumed to be transported on the next ship. In contrast to that, express containers are defined for urgent cargo and get prioritization on the ship they are booked for. This new segmentation has additional benefits for the shipper and the carrier. The shipper gains more control on the delivery speed and reliability of his shipment in the express segment and can profit from lower freight rates in the standard segment. The carrier's advantage is a higher freight rate for express containers, and in the standard segment the carrier achieves more flexibility in slot allocation, because of the option to postpone containers.

3 Slot Allocation Model

The optimization model provides the carrier with the best slot allocation for each segment and hence enables him to maximize the revenue of a ship. In this work, a path flow modeling approach is selected, as [2] show that the size and the solution speed of a path flow model outperforms the arc flow model. Similar models can be found in [2] or [4], but in contrast to these models, this slot allocation model includes the new segmentation with the possibility of container postponement.

The model includes a number of ports (OD) on the relevant loop and the legs ($(o,d) \in PP$ with $o,d \in OD$) connecting consecutive ports in the loop. The different paths in the network are the different possible routes ($b \in PA$) connecting different ports by one or more legs. The indicator $Y^1_{(o,d)b}$ is equal to 1, if the leg (o,d) is part of route b, and 0 otherwise. Other indicators are $Y^2_{(o,d)b}$ which is equal to 1, if the leg (o,d) is the start leg of the route and $Y^3_{(o,d)b}$ which is equal to 1, if the leg (o,d) is the end leg of the route. According to the new segmentation, the two time related segments ($s \in SE$) of express ($s = 1$) and standard ($s = 2$) containers are introduced. As container types ($t \in TY$), the dry and reefer types (TY_R) of 20' and 40' respectively and a 40' Dry High Cube are considered. In this model, the contractual container demand is treated as a constant because of the minimum quantity contracts. Hence, only the spot market containers are taken into account.

The model determines the number of slots (x^f_{stb}) that should be reserved for loaded (f) containers in segment s of type t on route b. The slot numbers can be seen as booking limits. The number of empty (e) containers (x^e_{tb}) which are repositioned is also considered for every route and container type. The third decision variable z^f_{tb} models the number of loaded standard containers that need to be postponed.

In the following, some more parameters and assumptions of the model are introduced. The slot capacity (CAP) in Twenty-foot Equivalent Units (TEU), the deadweight (DW) in tons and the number of reefer plugs (RP) of the ship is known. Each container type is assumed to have a known average weight (W^f_{tb} and W^e_t) and dimension (D_t). The dimensions for the different container types are measured in TEU with $D_t = 1$ for 20', $D_t = 2$ for 40' and $D_t = 2.25$ for 40' High Cube containers. For a reefer container, a plug is needed, because a loaded reefer container needs electronic power for refrigeration. The demand for empty (E^{in}_{td}) containers of every type at every port and for loaded containers (U_{stb}) of all types and segments on every route is forecasted. The number of empty containers (E^{out}_{to}) available for repositioning is also known for every port. An average freight rate (P_{stb}) for every segment, route and container type, and the average transportation costs for loaded (C^f_{tb}) and empty (C^e_{tb}) containers on every route are estimated.

Standard containers can wait for the next ship and get priority for it. Average costs for storing a container until the next departure (C^l_{tb}) are known for each type and route. Also the maximum storage capacity (SC_o), the number of postponed containers (A^f_{tb}) from the previous ship and the forecast of express containers for the next ship (V^f_{tb}) is known. Finally, the maximum number of express containers (G_o)

has to be restricted, because if delays occur the company needs some flexibility through standard containers to ensure their promised service for express containers.

The following model can be solved for a ship on the loop at the beginning of the booking period to determine the optimal booking limits for each segment. If the demand is different from the forecast, it is advisable to update the forecast and to reoptimize the model within the booking period.

$$max \sum_{s\in SE}\sum_{t\in TY}\sum_{b\in PA}(P_{stb}-C^f_{tb})x^f_{stb}+\sum_{t\in TY}\sum_{b\in PA}((P_{2tb}-C^f_{tb}-C^l_{tb})z^f_{tb}-C^e_{tb}x^e_{tb}) \quad (1)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^1_{(o,d)b}D_t(\sum_{s\in SE}x^f_{stb}+A^f_{tb}+x^e_{tb})\le CAP \qquad \forall(o,d)\in PP \quad (2)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^1_{(o,d)b}(W^f_{tb}(\sum_{s\in SE}x^f_{stb}+A^f_{tb})+W^e_t x^e_{tb})\le DW \qquad \forall(o,d)\in PP \quad (3)$$

$$\sum_{t\in TY_R}\sum_{b\in PA}Y^1_{(o,d)b}(\sum_{s\in SE}x^f_{stb}+A^f_{tb})\le RP \qquad \forall(o,d)\in PP \quad (4)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^2_{(o,d)b}(x^f_{1tb}+A^f_{tb})\le G_o \qquad \forall(o,d)\in PP \quad (5)$$

$$x^f_{1tb}\le U_{1tb} \qquad \forall t\in TY,\forall b\in PA \quad (6)$$

$$x^f_{2tb}+z^f_{tb}\le U_{2tb} \qquad \forall t\in TY,\forall b\in PA \quad (7)$$

$$\sum_{b\in PA}Y^3_{(o,d)b}x^e_{tb}\ge E^{in}_{td} \qquad \forall t\in TY,\forall(o,d)\in PP \quad (8)$$

$$\sum_{b\in PA}Y^2_{(o,d)b}x^e_{tb}\le E^{out}_{to} \qquad \forall t\in TY,\forall(o,d)\in PP \quad (9)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^1_{(o,d)b}D_t(z^f_{tb}+V^f_{tb})\le CAP \qquad \forall(o,d)\in PP \quad (10)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^1_{(o,d)b}W^f_{tb}(z^f_{tb}+V^f_{tb})\le DW \qquad \forall(o,d)\in PP \quad (11)$$

$$\sum_{t\in TY_R}\sum_{b\in PA}Y^1_{(o,d)b}(z^f_{tb}+V^f_{tb})\le RP \qquad \forall(o,d)\in PP \quad (12)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^2_{(o,d)b}(z^f_{tb}+V^f_{tb})\le G_o \qquad \forall(o,d)\in PP \quad (13)$$

$$\sum_{t\in TY}\sum_{b\in PA}Y^2_{(o,d)b}D_t z^f_{tb}\le SC_o \qquad \forall(o,d)\in PP \quad (14)$$

$$x^f_{stb}\ge 0 \text{ and integer} \qquad \forall s\in SE,\forall t\in TY,\forall b\in PA \quad (15)$$

$$x^e_{tb},z^f_{tb}\ge 0 \text{ and integer} \qquad \forall t\in TY,\forall b\in PA \quad (16)$$

The objective function 1 maximizes the sum of the expected revenue for the time related segments, the container types and the different routes. It includes empty container repositioning and storage costs for the standard containers which need to be stored until the next ship arrives. The constraints 2 and 3 deal with the capacity and the deadweight restrictions of the ship. The different dimensions of the container types and the average weights are taken into account. The loaded containers,

the empty containers and the postponed containers from the previous ship cannot exceed the total capacity or the deadweight of the ship. In equation 4, the reefer plug limitation is expressed. Constraint 5 takes care of the fact, that every port can only prioritize a specific number of express containers. The forecasted market demand in the different segments is modeled as an upper bound in the constraints 6 and 7. The equations 8 and 9 are for the empty container handling and ensure that enough empty containers are shipped into a port and that not more than the available containers are shipped out of a port. The constraints 10 to 13 are similar to the constraints 2 to 5 and model the restrictions for the next ship, including the forecasted demand for express containers. Equation 14 takes care of the fact that every port can store only a limited number of containers for the next ship. The final constraints 15 and 16 are the non-negativity and integrality constraints of the decision variables.

4 Conclusion

In this paper, a time related segmentation approach for the liner shipping industry is proposed and a slot allocation model based on this segmentation is introduced. This segmentation approach generates advantages for both parties involved. The new slot allocation model determines the revenue maximizing slot allocation strategy and helps the carrier to decide whether to accept or reject a booking request. The model takes into account the possibility of postponing a standard container to accept a more beneficial express container without losing the reputation on reliability and delivery speed. The power of defining the priority of a container is shifted from the carrier to the shipper, but in return the carrier can realize higher freight rates in the express container segment.

References

1. R. Hellermann. *Capacity Options for Revenue Management: Theory and Applications in the Air Cargo Industry*. Springer Berlin Heidelberg, 2006.
2. B. Løfstedt, D. Pisinger, and S. Spoorendonk. Liner shipping revenue management with repositioning of empty containers. Dept. of Computer Science – University of Copenhagen, 2008.
3. C. S. Lu. Evaluating key resources and capabilities for liner shipping services. *Transport Reviews*, 27(3): 285–310, 2007.
4. H.-A. Lu, C.-W. Chu, and P.-Y. Che. Seasonal slot allocation planning for a container liner shipping service. *Journal of Marine Science and Techology*, 18(1): 84–92, 2010.
5. L. Pompeo and T. Sapountzis. Freight expectations. *The McKinsey Quarterly*, 2: 90–99, 2002.
6. M. Stopford. *Maritime Economics*. Routledge New York, 2009.

E-Commerce Evaluation – Multi-Item Internet Shopping. Optimization and Heuristic Algorithms*

Jacek Błażewicz and Jędrzej Musiał

Abstract Report [11] states that 32% of EU customers make purchases in Internet stores. By 2013, almost half of Europeans are expected to make a purchase online, up from 21% in 2006. On-line shopping is one of key business activities offered over the Internet. However a high number of Internet shops makes it difficult for a customer to review manually all the available offers and select optimal outlets for shopping, especially if a customer wants to buy more than one product. A partial solution of this problem has been supported by software agents so-called price comparison sites. Unfortunately price comparison works only on a single product and if the customer's basket is composed of several products complete shopping list optimization needs to be done manually. Our present work is to define the problem (multiple-item shopping list over several shopping locations) in a formal way. The objective is to have all the shopping done at the minimum total expense. One should notice that dividing the original shopping list into several sub lists whose items will be delivered by different providers increases delivery costs. In the following sections a formal definition of the problem is given. Moreover a prove that problem is NP-hard in the strong sense was provided. It is also proven that it is not approximable in polynomial time. In the following section we demonstrate that shopping multiple items problem is solvable in polynomial time if the number of products to buy, n, or the number of shops, m, is a given constant.

Jacek Błażewicz (e-mail: Jacek.Blazewicz@cs.put.poznan.pl)
Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965 Poznan, Poland
Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Noskowskiego 12, 61-704 Poznan, Poland

Jędrzej Musiał (e-mail: Jedrzej.Musial@cs.put.poznan.pl)
Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965 Poznan, Poland

* The work was partially supported by the grant from the Ministry of Science and Higher Education of Poland.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_24,

We also described an heuristic algorithm we propose and try to find some connections to known and defined problems. The paper concludes with a summary of the results and suggestions for future research.

1 Introduction

On-line shopping is one of key business activities offered over the Internet. Report [11] states that 32% of EU customers make purchases in Internet stores. By 2013, almost half of Europeans are expected to make a purchase online, up from 21% in 2006. High number of Internet sellers provides strong competition and bring customers lower prices [6]. However comparison of all offers is harder for higher number of products, which differs not only in the price of an item, but also in shipping cost. If we would like to buy few items and make full investigation the complexity of offers comparison rise serious difficulty and rarely can be entirely performed manually. Using Price comparison sites applications could be helpful. Price comparison sites are a kind of software agents [2] designed to build a price rank for a single product among registered offers that fit to customer's query. It is worth noting that price ranking list built on-line on a customers' request expressed in a text query (product description) is a solution to a specific case of shopping, in which a customer wants to buy a single product. Multiple item shopping is not supported by price comparators available nowadays. As a result, price comparison sites play the role of recommender systems [10] which tend to detect a customers' preferences and interests in order to suggest products to buy.

Price sensitive discounts and shipping cost discounts are often used in Internet shops to attract customers. A typical example is the following advertisement in an Internet shop: "If the value of your purchase is at least 100 Euros, then your discount is 2%; 200 - 3%; 300 - 4%, 500 - 5%; 1000 - 10%". Many Internet shops offer free delivery if the price of a purchase exceeds a certain threshold.

In this paper we remind problem of optimal realization of multiple-item customer basket in Internet shops with zero delivery cost (section 2). In the following section 3 Internet shopping optimization problem (ISOP) was defined. In section 3.1 we defined interesting sub problem (specialization of the ISOP), which can be transformed into Facility Location Problem. Following section 4 describes heuristic algorithm for most complicated of the described ISOP cases. Last section concludes the topic and provides suggestions for future research.

2 Multiple-Item Shopping with Zero Delivery Cost

The first idea of shopping optimization was connected with self pick up - zero delivery cost (motivation comes from prescript medicines from pharmacy). A kind of software agent was created and described in [8]. The aim of the created algorithm

is to propose such a division of non-empty customer's shopping list into minimal number of baskets where each basket is fully realized in a single shop that realization of all baskets guarantees minimal cost of the entire shopping. In other words the logic of the mini-mini algorithm corresponds to the behavior of a thrifty consumer, who would like to buy all the products, pay as little as necessary, and, in the best case, do all his shopping in one location. Although the main optimization criteria is minimal cost of entire shopping realization, the mini-mini algorithm works in dialog mode with the customer and may ask if the customer can accept increasing number of shops to visit in order to save money. Savings potential is known on every step of the optimization process.

3 Multi-Item Internet Shopping

Let us define optimal realization of multiple-item customer basket in Internet shops as follows [1]. A single buyer is looking for a multiset of products $N = \{1,\ldots,n\}$ to buy in m shops. A multiset of available products N_l, a cost c_{jl} of each product $j \in N_l$, and a delivery cost d_l of any subset of the products from the shop to the buyer are associated with each shop l, $l = 1,\ldots,m$. It is assumed that $c_{jl} = \infty$ if $j \notin N_l$. The problem is to find a sequence of disjoint selections (or carts) of products $X = (X_1,\ldots,X_m)$, which we call a *cart sequence*, such that $X_l \subseteq N_l$, $l = 1,\ldots,m$, $\cup_{l=1}^{m} X_l = N$, and the total product and delivery cost, denoted as $F(X) := \sum_{l=1}^{m} \left(\delta(|X_l|)d_l + \sum_{j\in X_l} c_{jl}\right)$, is minimized. Here $|X_l|$ denotes the cardinality of the multiset X_l, and $\delta(x) = 0$ if $x = 0$ and $\delta(x) = 1$ if $x > 0$. We denote this problem as ISOP (Internet Shopping Optimization Problem), its optimal solution as X^*, and its optimal solution value as F^*.

The problem was proved to be NP-hard. In this context computational complexity justifies considering heuristic approaches as compromise solution balancing computational time and results close to optimum.

3.1 Flat Shipping Rate, No Price Discounts

Standard ISOP without price discounts and flat shipping rate, denoted as NO-DISCOUNTS, is equivalent to the well known FACILITY LOCATION PROBLEM (FLP). Main characteristics of the FLP are a space, a metric, given customer locations and given or not positions for facility locations. A traditional FLP is to open a number of facilities in arbitrary positions of the space (continuous problem) or in a subset of the given positions (discrete problem) and to assign customers to the opened facilities such that the sum of opening costs and costs related to the distances between customer locations and their corresponding facility locations is minimized. The equivalence of the traditional discrete FLP and problem NO-DISCOUNTS is

easy to see if customers are treated as products, positions for facility locations as shops, opening costs as delivery prices, and cost related to the distance between position i and customer j as the price p_{ij} of product j in shop i. Notice, however, that the general problem IS with price sensitive discounts cannot be treated as a traditional discrete FLP because there is no evident motivation for a discount on the cumulative cost in the sense of distances.

Discussions of FLPs can be found in [5, 4, 9, 7]. The traditional discrete FLP is NP-hard in the strong sense, so is the problem NO-DISCOUNTS. Solution approaches for discrete FLPs include Integer Programming formulations, Lagrangian relaxations, Linear Programming relaxations, Genetic Algorithms, graph-theoretical approaches, and Tabu Search, whose descriptions can by found in [9]. All these approaches can be used to solve problem NO-DISCOUNTS.

4 Heuristic Algorithm

Most interesting version of the optimization of multi-item Internet shopping problem is the one with pricewise linear discounts. Due to NP-hardness of the optimization problem a heuristic solution is proposed and evaluated for customer basket optimization problem to make it applicable for solving complex shopping cart optimization in on-line applications.

We developed simple heuristic algorithm which can work fast and provides better results than price comparison sites algorithms.

4.1 Algorithm Pseudocode Definition

Starting conditions with definition:

- $N = \{1, \ldots, n\}$ - products to buy.
- $M = \{1, \ldots, m\}$ - shops.
- p_{ij} - price for product j in shop i.
- rec_j - retailers recomended price for product j.
- d_i - non zero shipping cost for store i.
- $ds_i = d_i$ - shipping cost used for final computation.
- discounting function

$$f_i(P) = \begin{cases} P, & \text{if } 0 < P < 100, \\ 0.98P, & \text{if } P \geq 100, \end{cases}$$

  where P is the total standard price of books selected in store i.
- $sum = 0$ - overall cost of shopping.
- $\forall_{j \in N} R_j = 0$ - choosing indicator (realization list), for each product j.

Optimization procedure:

1. Let the products be ordered $1,\ldots,n$ where $rec_n \geq rec_{n-1} \geq rec_1$.
2. Select first product $j = 1$.
3. Select the shop i for product j such as $min(f_i(p_{ij}) + d_i)$; $i = 1,\ldots,m$.
4. $R_j = i$. $d_i = 0$;
5. Select next product $j = j + 1$. If $j \leq n$ goto 3.
6. $S = \sum_{i=1}^{M} f_i(P) + ds_i : d_i \neq 0$.
7. STOP.
   For better results algorithm could be re-run with different products order, such as $rec_n < rec_{n-1} < rec_1$, or random $rec_n ? rec_{n-1} ? rec_1$. Discounting function f_i could be build up to more complicated one.

5 Computational Experiment

A challenging step in experimental research was to create a model as close to real Internet shopping conditions as possible. We studied relationship between competitive structure, advertising, price and price dispersion over the Internet stores. Considering a group of representative products to be taken into account in computational experiment we chose books, because of their wide choice over the virtual (Internet) stores and popularity to buy over that kind of shopping channel. Creating model we took some information and computational results from [3]. It mainly focus on electronic bookstores model definition, prices, acceptances factor, retailer brand and, what is important for optimization problem model definition, price dispersion. One should also notice that consumers may choose among a large number of Internet bookstores.

Proposed algorithm was tested in close to real on-line market conditions. We built a model (including price of the products, shipping costs, etc.) according to live data available on the Internet. Algorithm optimization processing for all set of data (every of the set includes price range and flat shipping cost for every shop) were repeated many times to provide overall statistical results without outbound failures. For each pair (n,m) 10 instances were generated. In each instance, the following values were randomly generated for all i and j in the corresponding ranges. Delivery price: $d_i \in \{5,10,15,20,25,30\}$, *publisher's recommended price* of book j: $rec_j \in \{5,10,15,20,25\}$, and price of book j in bookstore i: $p_{ij} \in [a_{ij},b_{ij}]$, where $a_{ij} \geq 0.69r_j$, $b_{ij} \leq 1.47r_j$, and the structure of intervals $[a_{ij},b_{ij}]$ follow information in Table V in [3]. First computational tests show promising results. We perform over 800 tests. In the worst case, solution found by heuristic algorithm was less than 5% more expensive than the optimal solution, it was 38% cheaper than solutions provided by Price Comparison Sites without taking delivery prices into account and 5% cheaper than solutions provided by the upgraded algorithms of Price Comparison Sites which count delivery prices. The average values of the above mentioned deviations are 2%, 45% and 14%, respectively. These first good algorithmic results presented above is an initial step in studying and solving the multi-item shopping

problem. More detailed description and much wider analysis of the results would be the topic of an upcoming paper.

Summary

On-line shopping is one of key business activities offered over the Internet. In this note we remind shopping problem with zero delivery cost. Subsequently we defined the multiple-item shopping problem in a formal way. Moreover we developed and described an heuristic algorithm and try to find some connections to known and defined problems, such as Facility Location Problem. Modeling, computational complexity and algorithmic results presented in this paper is an initial step in studying and solving the multi-item Internet shopping problem. In the future, we will concentrate on the development of efficient exact and approximate solution procedures and efficiency verification through computer experiments as well as a worst-case analysis.

References

1. J. Błażewicz, M. Y. Kovalyov, J. Musiał, A. P. Urbański, and A. Wojciechowski. Internet Shopping Optimization Problem. *International Journal of Applied Mathematics and Computer Science*, 20(2): 385–390, 2010.
2. W. Chu,B. Choi, and M. R. Song. The Role of On-line Retailer Brand and Infomediary Reputation in Increasing Consumer Purchase Intention. *International Journal of Electronic Commerce*, 9(3): 115–127, 2005.
3. K. Clay, R. Krishnan, and E. Wolff. Prices and price dispersion on the Web: Evidence from the online book industry. Technical report, National Bureau of Economic Research, Inc., 2001.
4. H. A. Eiselt and C. L. Sandblom. *Decision analysis, location models, and scheduling problems*. Springer-Verlag, Berlin-Heidelberg-New York, 2004.
5. J. Krarup, D. Pisinger, and F. Plastriab. Discrete location problems with push-pull objectives. *Discrete Applied Mathematics*, 123: 363–378, 2002.
6. H. G. Lee. Do Electronic Marketplaces Lower the Prices of Goods? *Communications of the ACM*, 41(1): 73–80, 1998.
7. M. T. Melo, S. Nickel, and F. Saldanha-da-Gama. Facility location and supply chain management. *European Journal of Operational Research*, 196: 401–412, 2009.
8. J. Musiał and A. Wojciechowski. A customer assistance system: Optimizing basket cost. *Foundations of Computing and Decision Sciences*, 34(1): 59–69, 2009.
9. C. ReVelle, H. A. Eiselt, and M. Daskin. A bibliography for some fundamental problem categories in discrete location science. *European Journal of Operational Research*, 184: 817–848, 2008.
10. B. Satzger, M. Endres, and W. Kielssing. A Preference-Based Recommender System. In K. Baukhnecht, editor, *E-Commerce and Web Technologies*. Springer-Verlag, Berlin Heidelberg, 2006.
11. The Future Foundation. E-commerce across Europe – Progress and prospects. [online], 2008. http://www.eaca.be/_upload/documents/publications/E-commerce%20across%20Europe.pdf

I.6 Quantitative Models for Performance and Dependability (QMPD)

Chair: Pof. Dr. Markus Siegle (Universität der Bundeswehr München)

Mini-PC: Hans Daduna (Germany), Susanna Donatelli (Italy), William Knottenbelt (UK), Markus Siegle (Germany), Katinka Wolter (Germany)

In many ways, our society relies on the correct and timely functioning of large-scale distributed information and communication systems. We therefore solicit original contributions which address the model-based quantitative analysis of such concurrent systems, with the focus on performance, dependability, energy-efficiency, cost-optimal operation, vulnerability and safety. Systems of interest include parallel or distributed computing systems, web-based systems, production systems, logistics systems, wired or wireless communication systems, soft-real-time systems, safety-critical systems. The considered modelling formalisms are, among others, Markov chains, Markov Decision processes, Stochastic Petri Nets, Stochastic Process Algebra, (layered) queueing networks, fluid and hybrid stochastic models.

Both theoretical papers and practical case studies are called for. Reports on new modelling approaches, solution techniques, optimization methods, verification algorithms, data structures and tools are sought.

Modelling and Optimization of Cross-Media Production Processes

Pietro Piazzolla, Marco Gribaudo, Andrea Grosso, and Alberto Messina

Abstract In present days, the media industry is looking for concrete business opportunities in publishing the same product across different media. This requires careful planning to minimize the risks of each production without increasing its cost. We consider a cross-media production as a content that is explicitly produced to be distributed across multiple media. We formalize it as a collection of several Media Production Objects (MPOs): abstract representations of the production components such as audio and video footages, 3D models, screenplays, etc. that can be aggregated to become for example a film, a Video Game. We call a "View" the aggregated product ready to be published on a given channel. MPOs can have temporal constraints imposed to their realization, and are characterized by quality and costs metrics. We propose a model for the definition of cross media production, presenting an automatic procedure to create mixed-integer linear programming problems whose solution allows to optimize costs and qualities. This scenario opens a wide range of possibilities, that a MIP solver (in our case, $XPRESS-MP$) can handle as a black-box tool. This allows efficient what-if analysis under different scenarios.

1 Introduction

In present days, the media industry is starting to see concrete business opportunities in publishing the same product across different media [1]. No single production is worth the risk of investment if it has to be exploited only on one publication channel. This is because of a known tendency that sees the audience of a single channel

Pietro Piazzolla, Andrea Grosso
University of Torino, Torino, Italy. e-mail: {piazzolla.pietro, grosso}@di.unito.it

Marco Gribaudo
Politecnico di Milano, Milano, Italy. e-mail: gribaudo@elet.polimi.it

Alberto Messina
RAI - Centre for Research and Technological Innovation, Torino, Italy.e-mail: messina@rai.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_25,

diminishing because of the growth of new publication channels [5]. To minimize the risks of each production without increasing the costs related to production and advertising [1], and in order to maximize profits, the industry is compelled to find an efficient technological strategy for covering as many publication channels as possible yet in the production phases [5]. This is the field of cross-media productions.

Due to its industrial and non-scientific origin, the term cross-media lacks an exact and broadly accepted definition [6]. In the last decade the term has been used with different meanings depending on the context. In this paper we follow the definition given in [8], considering cross-media any content that is published almost at the same time across multiple media. Many works have been recently produced in this research area. For example, [2] and [3] consider intelligent analysis and synthesis of multimedia data as substantial enabling factors in making interactive, multi-channel, multi-purpose and value-returning productions. Aside from the many issues arising during a cross-medial production, a Company should exploit all the possible publication channels to gain an edge over competitors.

Through this paper we propose a model to help media production companies in planning such a complex investment. The goal is to improve the quality and profits of the whole production while keeping low the costs for each publication channel.

2 Modelling Framework

Formally, we define a *Media Production Specification* (*MPS*) as a tuple:

$$MPS = (M, L_T, m_0) \tag{1}$$

where M is a set of *Media Production Objects* (*MPO*s), L_T is a set of *Media Types* and $m_0 \in M$ is the *root* object that represents the entire production. The set of Media Types $L_T = \{root, l_1, l_2, \ldots\}$, is a set of labels that represents different types of *Media Objects* (*MO*s). *MO*s are a representation of the concrete elements that can be realized by the production, such as audio, video, edited video, text, and so on. The special type *root* is used to identify the entire production.

An *MPO* represents the act of producing an *MO*. It can be an independent process (e.g.: recording audio footage) or it may represent the aggregation of other *MO*s (e.g.: mixing audio tracks). In the latter case we call *input* the set of *MO*s needed to perform the task modelled by the *MPO* itself. For the realization of each *MO* in the *input*, the production may choose among different strategies. The main focus of our work is the optimization of this choice.

Formally an *MPO* $m_i \in M$ is a the tuple:

$$m_i = (t_i, D_i, C_i, \Gamma_i, \sigma(), q()). \tag{2}$$

$t_i \in L_T$, defines the type of MO produced by m_i.
D_i represents the *duration* of the production. We imagine that the production can start as soon as all the required components are available, and that it will end D_i from its beginning.

$C_i = \{a_1, \ldots, a_{N_i}\}$ represents the *input* of m_i. In particular it is a list composed of N_i elements. Each element a_j specifies a set of alternative ways to produce an MO. We add the special type $\top$ to include the possibility of making m_i without this element. Formally $a_j \in L_T \times 2^{M \cup \top}$ is a tuple $a_j = (t_j, A_j)$ where t_j represents the type of the MO produced and $A_j = \{m_k, \ldots\} \in 2^{M \cup \top}$ defines the set of alternative MPOs that can be used to generate component a_j. Figure 1 gives a visual representation of C_i. If $C_i = \emptyset$, the MPO m_i is called *leaf* and represents an elementary object that does

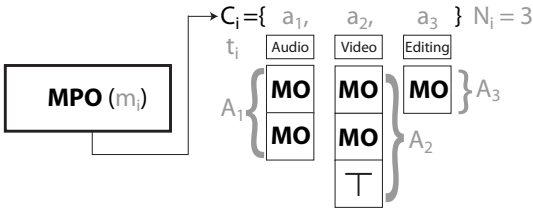


Fig. 1 The input of M_i: an example MPO.

not require other components to be produced.
$\Gamma_i = \{\gamma_1, \ldots, \gamma_{M_i}\}$ is a set of M_i production constraints. Each production constraint is used to define temporal characteristics that should be met by this particular production. We defined 3 types of constraints. Constraints of type $\{\text{begins,ends}\} \times \{\text{after,before}\} \times \tau$ are *absolute* constraints that state that the production should begin (end) before (after) τ. Constraints of the type $\{\text{begins,ends}\} \times \{\text{after,before}\} \times \tau \times \{\text{start,end}\} \times m$ are *relative* constraints and state that the production should begin (end) before (after) τ from the start (end) of production m. Finally, constraints of type $\{\text{ifAvailable}\} \times m$ specify that m can be used as an alternative only if it has already been produced as a component of some other MPO. This constraint is meant to force the reuse of already produced components, and to avoid the production of redundant elements.
$\sigma_i()$, and $q_i()$ are two functions that express respectively the *cost* and the *quality* of the production based on the chosen alternative. As for the quality measures $q_i()$, an additive for is assumed, i.e. $q_i(m_j) = Q_j + \sum_k q_i(m_k)$ where k ranges over the selected objects composing m_j.

3 Model Development

As far as the task of coordinating activities while sharing resources, time and/or resource constraints while optimizing some type of quality or cost criterion is concerned, the problem is easily rewritable as a kind of (project) scheduling problem. Particularly, the scheduling problem can be written as a MIP problem and handled by a "black-box" MIP solver (say CPLEX, or XPRESS-MP); this admittedly creates a concern about the size of the instances that can be tackled, but the MIP-solvers technology has known dramatic advances in recent years.

Variables of the Model. The basic variables of the model are a set of binary variables $x_i \in \{0,1\}$, $i \in M$, where $x_i = 1$ iff object $i \in M$ is realized. Binary variables $y_{ik} \in \{0,1\}$ are introduced in order to model the following choice: a MO m_k is used to build a MO m_i iff $y_{ik} = 1$ — note that not every (i,k) pair needs to be used. For every y_{ik} used in the model we also introduce a nonnegative auxiliary variable $\bar{z}_{ik} \geq 0$. A variable $z_i \geq 0$ is associated with the selected quality measure of MO m_i.

Modeling the Quality. Part of the model is concerned with deciding which MO are to be produced in order to consistently create the root object m_0 while maximizing its quality. The model objective is to maximize z_0 subject to the following relevant constraints, written for every m_i.

Let $m_i = (t_i, D_i, C_i, \Gamma_i, \sigma(), q())$ accordingly with equation (2). We visit each m_i and the corresponding C_i, performing the following actions.

We visit the set of MO starting with the root m_0 and write the following constraint sets.

1. Write for each $(t_j, A_j) \in C_i$ such that $\top \notin A_j$:
$$\sum_{k \in A_j} y_{ik} = 1. \tag{3}$$
That is, one element of A_j must be selected for building m_i.
2. Write for each $(t_j, A_j) \in C_i$ such that $\top \in A_j$:
$$\sum_{k \in A_j} y_{ik} \leq 1. \tag{4}$$
That is, at most one (but possibly none) object from A_j must be selected.
3. Write for each $(t_j, A_j) \in C_i$:
$$\bar{z}_{ik} \leq z_k + M(1 - y_{ik}) \qquad \text{for all } m_k \in A_j, \tag{5}$$
$$\bar{z}_{ik} \leq M y_{ik} \tag{6}$$
$$z_i \leq Q_i + \sum_{j:\, A_j \in C_i} \sum_{k:\, m_k \in A_j} \bar{z}_{ik}. \tag{7}$$
This set of constraints fixes z_i to the quality value computed as an additive function of its components' qualities.
4. Write for each m_i, m_k:
$$\sum_k y_{ik} \leq x_k. \tag{8}$$
That is, if a MO m_k is used to build m_i, m_k must also be produced.

Other Constraints. Temporal precedences are taken into account by means of precedence constraints; if t_i is a real variable representing the starting time of (the implementation process) of MO $m_i \in M$, for each pair $i, j \in M$ such that i must be completed before j starts we add the constraint $t_i + D_i \leq t_j + M(1 - x_i)$, where the big-M part allows for avoiding to implement i.

Finally, the presence of a given budget B is easily taken into account by an additional constraint $\sum_{i \in M} \sigma_i x_i \leq B$.

4 Results – Model Exploitation

As an example of our model we propose the Media Production Specification in Figure 2. It is a formalization of a simplified yet plausible production aimed to produce a book, its movie, and maybe its video game. The production also has a supporting website that can or not have its community centered around a forum. In Figure 2 the *Type* of the *MPO*s is indicated in the upper part, while the lower

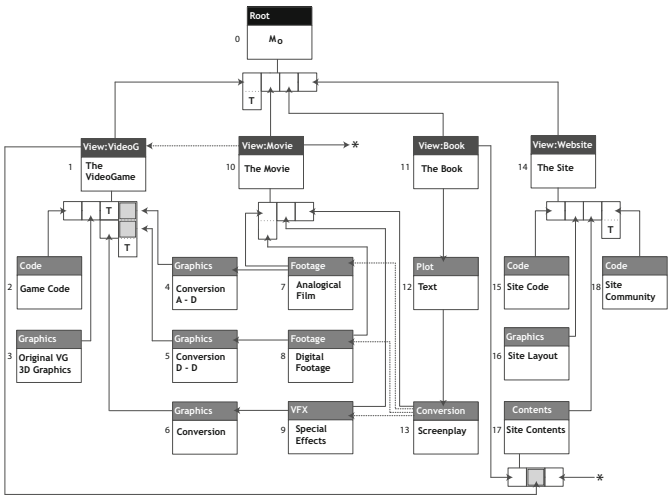


Fig. 2 A simple *MPS*.

part contains its name. Each connected box represents one of the *MPO*'s required composing elements, and a gray box represents an $ifAvailable$ constraint. When *MPO*s require just one component to be completed, the grid of boxes is omitted. The dotted arcs represent temporal constraints.

The model is composed by nineteen elements representing the Video Game, The Movie, The Book, The Site and their components. We established a single cost metric for the *MPO*s: $L_c = (Euro)$. For the quality L_q we chose two metrics: *Dissemination Quality* (q_D) and *Experience Quality* (q_E) so that $L_q = \{q_D, q_E\}$. The first one measures the advantages gained by the production in using the more possible media channels. The second one is the measure of the matching between the *MPO* produced and the expectancies/needs the Audience has about it. Equipped with realistic data, the model is translated into a MILP model that is easily solved by XPRESS-MP[1] within few seconds.

[1] XPRESS-MP provided by Fair Isaac Corporation under academic partnership.

We worked out the example in Figure 2 with the following data for each object.

ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Cost (10 keuro)	10	8	24	24	4	2	3	100	80	6	50	1.5	0.5	4	0.4	0.3	0.6	0.2	0.2
Dur. (weeks)	16	9	15	25	2	2	5	6	6	4	32	52	10	8	1	1	4	1	3
Q_E[2]	1	4	2	3	2	2	2	4	2	2	3	3	3	2	2	2	2	2	4
Q_D[2]	4	4	1	2	2	1	1	2	2	2	3	4	5	1	4	0.5	1	3	3

Solving the model for two budget scenarios with $B = 2000$ and $B = 3000$ keuro and optimizing for different metrics $z_0 = q_E(m_0)$ and $z_0 = q_D(m_0)$ respectively we got different configurations for the global production.

If we consider the optimization of $q_E(m_0)$, with the high budget, we obtain that $M = \{0,1,2,3,5,6,8,9,10,11,12,13,14,15,16,17,18\}$, and a total $q_E(m_0) = 147$. If the budget is the lower one, we obtain $M = \{0,8,9,10,11,12,13,14,15,16,17,18\}$ and a total $q_E(m_0) = 81$. If we optimize instead $q_D(m_0)$ the maximum are 163.5 and 97.5 for the two budget amounts respectively.

The computation time needed to solve the model — on a PC with Athlon 2.6 GHz processor and 512 MB RAM — is practically negligible, hence suggesting that larger models can be handled.

We are currently investigating other developments for these models and techniques, involving also a complete generation of the efficient frontier with two or more quality criteria, and generating robust solutions under different scenarios.

References

1. E. Aarseth. The Culture and Business of Cross-media Productions. *Popular Communications*, 4(3): 202–211, 2006.
2. S. M. Bettega, F. Fioravanti, L. Gigli, G. Grassi, and M. B. Spinu. Automated solutions for Cross Media Content and Multi-channel Distribution. In *Proc. of AXMEDIS 2008*, pages 57–62, 2008.
3. M. Czyrnek, E. Kusmierek, C. Mazurek, and M. Stroinski. New services for iTVP content providers to manage live and on-demand content streaming. In *Proc. of AXMEDIS 2008*, pages 180–186, 2008.
4. A. Elberse and J. Eliashberg. Demand and supply dynamics for sequentially related products in international markets: The case of motion pictures. *Marketing Science*, 22(3): 329–354, 2003.
5. J. Lemke. Critical Analysis across Media: Games, Franchises, and the New Cultural Order. In *Proc. of I Int. Conf. on Critical Discourse Analysis*, 2004.
6. K. SandKuhl. First Steps to Cross Media Publishing and Multimodal Documents. In *Principles of Document Processing*, 1997.
7. A. De Vany. *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry*. Routledge, London, 2004.
8. A. A. Veglis. Modelling Cross Media Publishing. In *Proc. of III Int. Conf. on Internet and Web Applications and Services*, 2008.

[2] Q_E and Q_D are here estimated in the discrete range $\{0,\ldots,5\}$, being 5 the maximum possible quality contribute added to the production by a given *MO* itself, before adding its components' quality. A more precise and realistic estimation is possible but outside the scopes of this paper. See [4] or [7], among others, for more details about word-of-mouth and audience calculation.

Diffusion Approximation for a Web-Server System with Proxy Servers

Yoshitaka Takahashi, Yoshiaki Shikata, and Andreas Frey

Abstract It is an important and urgent Operations Research (OR) issue to evaluate the delay in a web-server system handling internet commerce real-time services. Usually, proxy servers in differently-located sites enable us to shorten the web-server access delay in order to guarantee the quality of real-time application services. However, there exists almost no literature on the queueing analyses for the web-server system with proxy servers. The goal of this paper is to provide a queueing analysis for the web-server system. We derive the statistics of the individual output processes from the proxy servers. Regarding the unfinished workload in the web-server system with input as a diffusion process, we derive a mean-delay explicit formula.

1 Introduction

A tremendous increase in WWW (world wide web) service users frequently causes insufficient communication facilities and resources. Very recently multimedia contents become common more and more, and the internet traffic monotonically and rapidly increases due to ordinary characters (texts), codes, voices, and moving pictures. This increasing traffic leads to the deterioration of web-access response times. The response time deterioration should be promptly solved because of some web

Yoshitaka Takahashi
Faculty of Commerce, Waseda University, 169-8050 Tokyo, Japan, e-mail: yoshitak@waseda.jp

Yoshiaki Shikata
Faculty of Informatics for Arts, Shobi University, 350-1153 Saitama, Japan, e-mail: y-shikata@shobi-u.ac.jp

Andreas Frey
Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrueck, 49076 Osnabrueck, Germany, e-mail: a.frey@hs-osnabrueck.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_26,

applications for a real-time commerce. Queueing analysis of web-server access operation is requested to guarantee the quality of real-time services.

The main goal of this paper is to provide a queueing modeling and analysis for the web-server system with proxy servers.

2 Queueing Modeling for Web-Server System

Partly generalizing the model in [10], we make the following stochastic assumptions for the web-server system with N proxy servers:
a) For any $i = 1,2,\ldots,N$, the customer-arrival process for the i-th proxy server forms a renewal process with independent and identically distributed (iid) inter-arrival time, A_i, with arrival rate λ_i, and squared coefficient of variation as

$$E(A_i) = \frac{1}{\lambda_i}, \qquad C_{A_i}^2 = \frac{V(A_i)}{E(A_i)^2} \qquad (i = 1,2,\ldots,N)$$

Here, a customer corresponds to the client who requests the contents (an object) to the proxy server.
b) With probability p_i, an arriving customer leaves the system. This corresponds to the situation where the proxy server finds the client's requesting contents (object) within itself (the proxy cache).
c) With probability $1 - p_i$, an arriving customer enters the infinite-capacity single-server queueing system. Here, the queueing discipline is assumed to be FCFS (First-Come-First-Served); see [10] for a practical application. This corresponds to the situation where the proxy server cannot find the client's requesting contents within itself and the proxy server makes the client enter the wait-state to transmit the request to the web-server (single-server).
d) The service time is iid with mean and squared coefficient of variation as

$$E(B) = \frac{1}{\mu}, \qquad C_B^2 = \frac{V(B)}{E(B)^2}$$

Here, the service time B corresponds to the time period from the epoch at which the proxy server receives a client's request which is not found within the proxy server until the epoch at which the proxy server obtains the contents (the object). Let $B(x)$ be the cumulative distribution function of the service time, $B(x) = P(B \leq x)$.
Remark: Let f_k be the frequency where a web-access request has popularity degree k ($k = 1,2,\ldots,D$). Letting D denote the total number of contents, the Zipf's law is widely applied to obtain

$$f_k = \frac{k^s}{\sum_{j=1}^{D} j^s} \qquad (k = 1,2,\ldots,D)$$

for some s ($s > 0$). In this situation, we have

$$p_i = \sum_{k=1}^{K_i} f_k = \frac{\sum_{k=1}^{K_i} k^s}{\sum_{j=1}^{D} j^s} \qquad (i = 1,2,\ldots,N)$$

where K_i denotes the contents capacity (cache size) of the i-th proxy server.

3 The Input Process to Web-Server System

Note that the input process to the web-server system is the superposition of the output processes from N proxy servers. For any i ($i = 1,2,\ldots,N$), let $A_i^*(s)$ be the Laplace-Stieltjes Transform (LST) of the inter-arrival time distribution of A_i. Consider the inter-arrival time, A_{is}, of the substantial customers who will enter to the web-server system. Since the number of times where the i-th proxy server cannot find the requesting contents within itself follows a geometric distribution with parameter p_i, we have the LST of the inter-arrival time distribution of the substantial customers as

$$A_{is}^*(s) = \frac{(1-p_i)A_i^*(s)}{1-p_i A_i^*(s)} \qquad (i = 1,2,\ldots,N)$$

The first two moments are obtained by taking the derivatives of the LST above at $s = 0$. So, the squared coefficient variation of A_{is} is given by

$$C_{is}^2 \equiv \frac{V(A_{is})}{E(A_{is})^2} = (1-p_i)C_{Ai}^2 + p_i \qquad (i = 1,2,\ldots,N)$$

4 Diffusion Approximation

Let $\{V(t)\}$ be the unfinished work process in the single-server FCFS system, and let $f(x,t)$ be the probability density function (pdf) of $V(t)$. When approximating the unfinished work by a diffusion process, we need boundary conditions. The reflecting barrier (RB) and elementary return (ER) boundaries are well used in the queueing systems. We consider both boundaries as follows.

4.1 Diffusion Equation with RB ([4], [6], [7], [8])

If we set the reflecting barrier (RB) at space origin ($x = 0$), we assume the pdf $f(x,t)$ satisfies:

$$\frac{\partial f}{\partial t} = -\alpha \frac{\partial f}{\partial x} + \frac{\beta}{2}\frac{\partial^2 f}{\partial x^2}, \qquad 0 = \left[-\alpha f + \frac{\beta}{2}\frac{\partial f}{\partial x}\right]_{x=0}, \qquad \lim_{x\to\infty} f(x,t) = 0$$

4.2 Diffusion Equation with ER ([2], [3], [5], [9])

If we set the elementary return (ER) boundary at space origin ($x = 0$), we assume the pdf $f(x,t)$ satisfies:

$$\frac{\partial f}{\partial t} = -\alpha \frac{\partial f}{\partial x} + \frac{\beta}{2} \frac{\partial^2 f}{\partial x^2} + \sum_{i=1}^{N} \lambda_{is} \pi_0(t) \frac{dB(x)}{dx}$$

$$\frac{d\pi_0(t)}{dt} = -\sum_{i=1}^{N} \lambda_{is} \pi_0(t) + \left[-\alpha f + \frac{\beta}{2} \frac{\partial f}{\partial x} \right]_{x=0}, \qquad \lim_{x \to 0} f(x,t) = \lim_{x \to \infty} f(x,t) = 0$$

Here, the parameters (α, β) are called as the infinitesimal mean and variance, and obtained in [2] and [9] as

$$\alpha = \sum_{i=1}^{N} \frac{\lambda_{is}}{\mu} - 1, \qquad \beta = \sum_{i=1}^{N} \frac{\lambda_{is}}{\mu^2} (C_{is}^2 + C_B^2)$$

The steady-state solution pdf $f(x)$ of the diffusion equations are obtained in [4] under RB boundary, and in [2] (corrected in [9] under ER boundary, as
(RB solution)

$$f(x) = -\frac{2\alpha}{\beta} \mathrm{e}^{\frac{2\alpha}{\beta} x}$$

from which we have the mean unfinished work

$$E(V) = \int_0^\infty x f(x) dx = -\frac{\beta}{2\alpha}$$

(ER solution)

$$f(x) = \frac{2\pi_0}{\beta} \sum_{i=1}^{N} \lambda_{is} \int_0^\infty (1 - B(y)) \mathrm{e}^{-\frac{2\alpha y}{\beta}} dy \cdot \mathrm{e}^{\frac{2\alpha x}{\beta}}$$

$$\pi_0 = -\alpha$$

from which we have the mean unfinished work as in [9]

$$E(V) = \int_0^\infty x f(x) dx = \frac{1}{2} \sum_{i=1}^{N} \lambda_{is} \left[E(B^2) - \frac{\beta}{\alpha} \frac{1}{\mu} \right]$$

Thus, we can evaluate the mean unfinished work explicitly.

The mean unfinished work $E(V)$ is the time average, while the mean waiting time $E(W)$ is the customer average. Thus, Brumelle's law [1] is applied to find the relationship between $E(V)$ and $E(W)$ as

$$E(W) = \frac{E(V) - \frac{\sum_{i=1}^{N} \lambda_{is}}{2} E(B^2)}{\sum_{i=1}^{N} \frac{\lambda_{is}}{\mu}}$$

Note that [2] has not used this relationship but regarded $E(W)$ as $E(V)$. Little's law [6] yields the mean number of customers in the system, $E(N)$ as

$$E(N) = \sum_{i=1}^{N} \lambda_{is} E(W) + \sum_{i=1}^{N} \frac{\lambda_{is}}{\mu}$$

We finally obtain the mean response time at the i-th proxy server as

$$E(D_i) = (1 - \rho_i)[E(W) + \frac{1}{\mu}]$$

Remarks:
(a) For Poisson-input ($C_{Ai} = 1$) system with single proxy server ($N = 1$), our approximation is reduced to

$$E(V) = E(W) = \frac{\lambda(1-p)E(B^2)}{2\left[1 - \frac{\lambda(1-p)}{\mu}\right]}$$

with omitting the index i indicating the number of proxy server, which is seen to be consistent with the exact result obtained in [10]. Also, the PASTA (Poisson Arrivals See Time Averages [11]) property is seen to be valid for our unfinished-workload based diffusion approximations with both RB and ER boundaries.
(b) For the $E_2/E_2/1$ system with single proxy server ($N = 1$), we show in Figure 2 our approximations for the mean number of customers as well as simulation results with 95 % confidence interval as dots.

5 Conclusion

We have modeled and analyzed the web-server system with proxy servers via the diffusion approximation. We have shown that our approximate formula is consistent with the previously-obtained exact result in [10]. As for a further study it would remain to investigate the retrial-input queueing system where the web users click a reload button impatiently when they experience some delay.

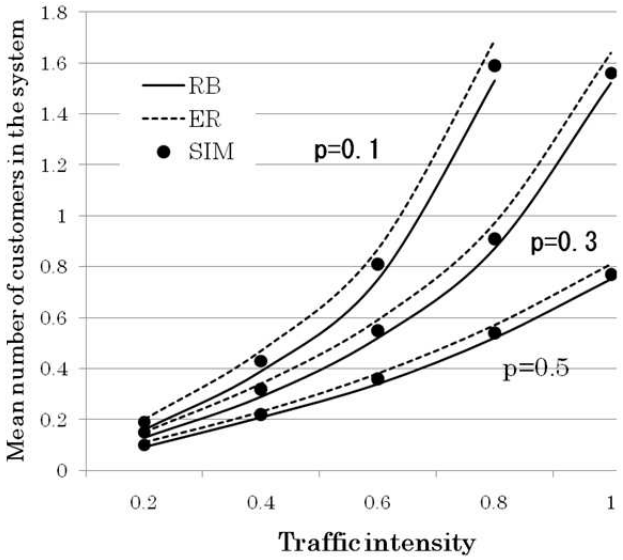


Fig. 1 The mean number of customers in the two-stage Erlang input and service time ($E_2/E_2/1$) system as a function of ρ with probability p that a client request finds its contents in the proxy cache.

References

1. S.L. Brumelle. On the relationship between customer and time averages in queues. *J. Appl. Prob.*, 8(3): 508–520, 1971.
2. E. Gelenbe. Probabilistic models of computer systems, Part II, Diffusion approximations, waiting times and batch arrivals. *Acta Informatica*, 12: 285–303, 1979.
3. E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. Academic Press, New York, 1980.
4. D.P. Heyman. A diffusion model approximation for the GI/G/1 queue in heavy traffic. *Bell System Tech. J.*, 54: 1637–1646, 1975.
5. T. Kimura. A unified diffusion model for the state-dependent queues. *Optimization*, 18: 265–283, 1987.
6. L. Kleinrock. *Queueing Systems, Volume II: Computer Applications*. John Wiley and Sons, New York, 1976.
7. H. Kobayashi. Applications of the diffusion approximations to queueing networks, I: equilibrium queue distributions. *J. ACM*, 21: 316–328, 1974.
8. G.F. Newell. *Applications of Queueing Theory*. Chapman & Hall, London, 1982.
9. Y. Takahashi. Diffusion approximation for the single-server system with batch arrivals of multi-class calls. *IECE Transactions*, J96-A(3): 317–324, 1986.
10. Y. Takahashi and T. Takenaka. Performance modeling a web-server access operation with proxy-server caching mechanism. In *Proceedings of the 45-th ORSJ (Operations Research Society of Japan) Symposium*, pages 1–9, 2001.
11. R.W. Wolff. Poisson arrivals see time averages. *Operations Res.*, 30: 223–231, 1982.

Large-Scale Modelling with the PEPA Eclipse Plug-In

Mirco Tribastone and Stephen Gilmore

Abstract We report on recent advances in the development of the *PEPA Eclipse Plug-in*, a software tool which supports a complete modelling workflow for the stochastic process algebra PEPA. The most notable improvements regard the implementation of the *population-based* semantics, which constitutes the basis for the aggregation of models for large state spaces. Analysis is supported either via an efficient stochastic simulation algorithm or through *fluid approximation* based on ordinary differential equations. In either case, the functionality is provided by a common graphical interface, which presents the user with a number of *wizards* that ease the specification of typical performance measures such as average response time or throughput. Behind the scenes, the engine for stochastic simulation has been extended in order to support both transient and steady-state simulation and to calculate confidence levels and correlations without resorting to external tools.

1 Introduction

The stochastic process algebra PEPA [5] has a long history of software tool support [4, 2, 1]. The *PEPA Eclipse Plug-in* has been recently proposed as an implementation of the language which supports the whole model development process with a rich graphical interface. References [10, 8] provide an overview of the main plug-in features such as static analysis, numerical solution of the underlying continuous-time Markov chain (CTMC), and the graph visualisation toolkit.

Mirco Tribastone
Institut für Informatik, Ludwig-Maximilians-Universität München, Germany,
e-mail: tribastone@pst.ifi.lmu.de

Stephen Gilmore
School of Informatics, The University of Edinburgh, Scotland, e-mail: stg@inf.ed.ac.uk

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_27,

```
/* Rate declarations */
r = 1.0;
s = 4.5;
t = 5.5;
/* Sequential component Process */
Process1 = (use, r).Process2;
Process2 = (think, s).Process1;
/* Sequential component CPU */
CPU1 = (use, r).CPU2;
CPU2 = (reset, t).CPU1;
/* System equation */
Process1[8] <use> CPU1[4]
```

Fig. 1 A sample PEPA model with two sequential processes. The component `Process` interposes some thinking time between uses of a `CPU`, modelled as a synchronisation over the action type `use`. The `CPU` performs some reset operation, e.g., a context switch after each use. The system equation considers eight processes and four CPUs in total. In this model, a process may use any of the available CPUs.

The present paper is concerned with a recently added module which implements the *population-based* semantics for PEPA [9].[1]

Let us re-use the simple PEPA model presented in [8], shown here in Figure 1, to give a brief and informal introduction to this semantics, which is preparatory for the discussion on the tool implementation. The state representation in this interpretation is a vector of counting variables, each representing the number of components which exhibit a specific local behaviour. In the running example, the state vector may be denoted as follows:

$$\boldsymbol{\xi} = (\xi_{Process_1}, \xi_{Process_2}, \xi_{CPU_1}, \xi_{CPU_2}),$$

and hence the initial state is $(8,0,4,0)$. The population-based semantics gives rise to *generating functions*, denoted by $f_\alpha(\xi, l)$, which give the rate at which an activity of type α is executed, and the change to a state due to its execution through the vector l. For instance, the shared action `use` is captured by the function

$$f_{use}\big(\xi, (-1,1,-1,1)\big) = \min(r\,\xi_{Process_1}, r\,\xi_{CPU_1}),$$

which says that `use` decreases the population counts of `Process1` and `CPU1` and, correspondingly, increases the population counts of `Process2` and `CPU2` at a rate which is dependent upon the current state. For instance, the following transition in the CTMC for the initial state may be then inferred:

$$(8,0,4,0) \xrightarrow{\min(1.0\times 8.0, 1.0\times 4.0)} (8,0,4,0) + (-1,1,-1,1) \equiv (7,1,3,1).$$

Extracting generating functions from a PEPA model presents little computational challenge because it does not require the exploration of the whole state space of the CTMC. In the plug-in, the *Differential Analysis View* updates the generating functions of the currently active model whenever its contents are saved. Figure 2 shows a screen-shot with the three generating functions of our running example.

[1] The latest release of the PEPA Eclipse Plug-in is available for download at http://www.dcs.ed.ac.uk/pepa/tools/plugin/.

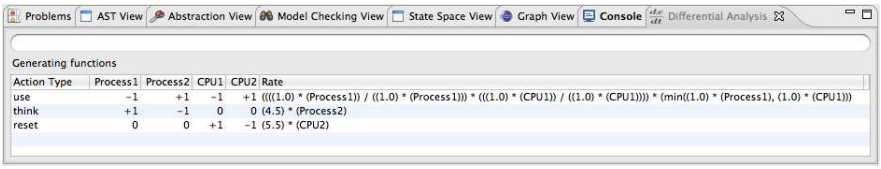


Fig. 2 Differential Analysis View for the model in Figure 1.

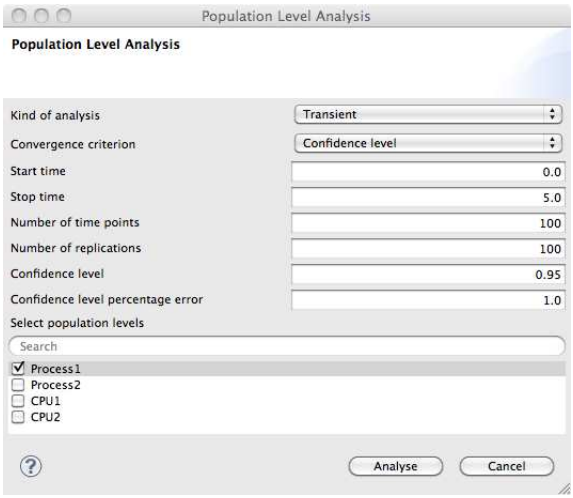


Fig. 3 Stochastic simulation dialogue box.

2 Model Analysis

Stochastic simulation. The generating functions contain all the necessary information for model analysis. They admit a straightforward stochastic simulation algorithm. Given a state $\hat{\xi}$, the evaluation of each of the generating functions $f_\alpha(\hat{\xi}, l)$ of the model gives the relative probabilities with which each action may be performed. Drawing a random number from the resulting probability density function decides which action is to be taken, and thus the corresponding target state $\hat{\xi} + l$. This procedure may be repeated until conditions of termination of the simulation algorithm are met. The PEPA Eclipse Plug-in implements transient and steady-state simulation algorithms (see Figure 3). Transient simulation is based on the method of independent replications, cfr., e.g., [7]. The user is required to select which components of the state vector are to be kept track of, the start and stop time of the simulation, the number of equally spaced time points of interest over the simulation interval, the maximum number of replications allowed, and a desired confidence level. Two stopping criteria may be used. The confidence-level criterion terminates the simulation when the user-specified confidence interval is within a given percentage of the statistical mean. If convergence is not reached within the maximum number of

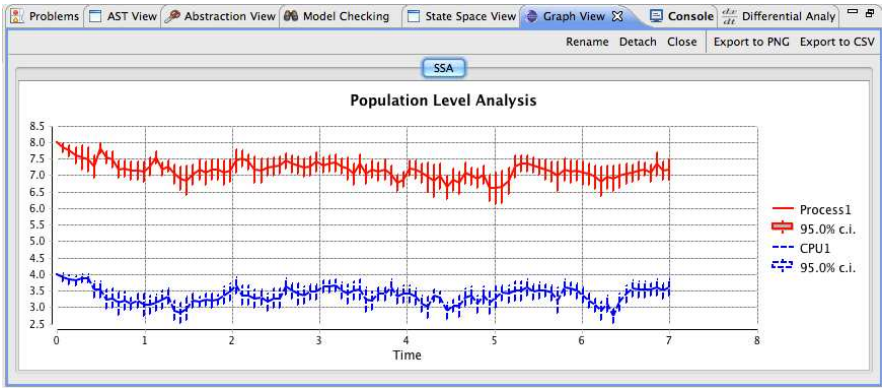


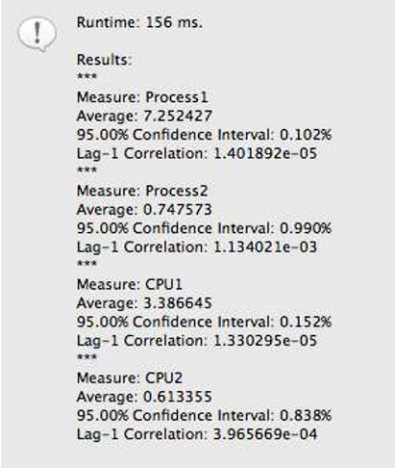
Fig. 4 Results of a transient stochastic simulation.

replications allowed, the results are reported together with a warning. Alternatively, the simulation may be simply stopped when the maximum number of replications is reached. In either case, the tool presents the results in a graph with error bars giving the confidence levels obtained (see Figure 4).

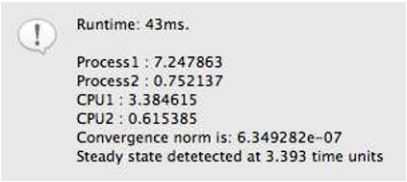
Steady-state simulation is performed with the method of *batch means*. Using a similar interface to that shown in Figure 3, the user is presented with a dialogue box to set up the following parameters: length of the transient period, during which samples are discarded; confidence level desired; maximum number of batches. The duration of each batch is set to ten times the transient period.[2] At the end of a batch, the algorithm checks whether the tracked population counts have reached the desired confidence level. If the maximum number of batches is reached, the algorithm returns with a warning of potentially bad accuracy. The results are presented in a dialogue box which gives the average steady-state value and the confidence level for each tracked population count (see Figure 5(a)). The lag-1 correlation is also computed as an indicator of statistical independence between adjacent batches [7].

Differential Analysis. The vector $x(t)$ of functions which is solution of the ordinary differential equation (ODE) $\frac{dx(t)}{dt} = \sum_l \sum_\alpha l f_\alpha(x(t), l)$ provides an approximating continuous trajectory of the population counts of the PEPA components as a function of time. The plug-in supports the numerical integration of the initial value problem corresponding to the model equation of the PEPA description. In the running example, the initial state of the vector $x(t) = (x_{Process_1}(t), x_{Process_2}(t), x_{CPU_1}(t), x_{CPU_2}(t))$ is $x(0) = (8,0,4,0)$. The plug-in supports both transient and steady-state analysis. In either case, the user is required to specify the integration interval and the absolute and relative tolerance errors for the numerical solver, based on the `odetojava` library [6]. For transient analysis, a mesh of time points of interest must be specified using the triple `start time`, `time step`, `stop time`. At those points the solution of the ODE will be computed and the results will be returned in the form of a graph. For steady-state analysis, the user is also required

[2] This parameter is currently not modifiable. Future releases will expose this setting to the user.



(a) Stochastic Simulation



(b) ODE Analysis

Fig. 5 Steady state results displayed in dialogue boxes.

to specify a tolerance error for equilibrium detection. At each time point τ analysed during the numerical integration, if the Euclidean norm of $\frac{dx(t)}{dt}|_{t=\tau}$ is below that threshold, the model is said to have reached equilibrium, and the corresponding solution vector $x(\tau)$ is returned. The user interface for either form of differential analysis is structurally similar to that for stochastic simulation—the top area of a dialogue box permits the set-up of the numerical integration, whereas in the bottom area the user chooses the components to visualise. As with steady-state stochastic simulation, the results are presented in a dialogue box (see Figure 5b).

Throughput Calculation. The generating functions can be interpreted as giving the throughput of an action type in each state ξ of the Markov chain. For instance, if the simulation model collects samples of $\min(r\,\xi_{Process_1}, r\,\xi_{CPU_1})$ then the statistics of this random variable will give the throughput of `use`. The graphical interface of Figure 3 is partially re-used to specify the simulation settings. However, in the bottom area are listed all the action types declared in the PEPA model. A fluid approximation of throughput, which is also implemented, is given by the evaluation of the function $f_\alpha\big(x(t), l\big)$. Finally, throughput and population-count measures can be combined in order to compute average response times, using an approach which was presented in [3]. For further information on the computation of these performance indices, the reader is referred to [11].

3 Conclusion

The current state of the PEPA Eclipse plug-in presented in this paper is a major improvement with respect to the range of evaluation tools for large-scale models. The modeller has now complete control of the confidence levels for transient and steady-state stochastic simulation. A unified user interface with fluid approximation allows seamless switching between these two techniques. As far as future work is concerned, the most pressing feature yet to be implemented is a framework to support *what-if* analysis, which is currently available only for models solvable by numerical solution of the equilibrium probability distribution of the underlying CTMC.

Acknowledgements This work has been partially carried out while M.T. was with the School of Informatics at the University of Edinburgh.

References

1. J.T. Bradley, N.J. Dingle, S.T. Gilmore, and W.J. Knottenbelt. Extracting passage times from PEPA models with the HYDRA tool: A case study. In S. Jarvis, editor, *Proceedings of the Nineteenth annual UK Performance Engineering Workshop*, pages 79–90, University of Warwick, July 2003.
2. Allan Clark. The ipclib PEPA Library. In *Fourth International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 55–56, Edinburgh, Scotland, UK, 17–19 September 2007. IEEE Computer Society.
3. Allan Clark, Adam Duguid, Stephen Gilmore, and Mirco Tribastone. Partial Evaluation of PEPA Models for Fluid-Flow Analysis. In *EPEW*, volume 5261 of *LNCS*, pages 2–16, 2008.
4. S. Gilmore and J. Hillston. The PEPA Workbench: A Tool to Support a Process Algebra-based Approach to Performance Modelling. In *Proceedings of the Seventh International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, number 794 in Lecture Notes in Computer Science, pages 353–368, Vienna, May 1994. Springer-Verlag.
5. J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
6. M. Patterson and J. Spiteri. odeToJava Library. http://www.netlib.org/ode/odeToJava.tgz, 2002.
7. William J. Stewart. *Probability, Markov Chains, Queues, and Simulation*. Princeton University Press, 2009.
8. M. Tribastone, A. Duguid, and S. Gilmore. The PEPA Eclipse Plug-in. *Performance Evaluation Review*, 36(4): 28–33, March 2009.
9. M. Tribastone, S. Gilmore, and J. Hillston. Scalable Differential Analysis of Process Algebra Models, 2010. *Transactions on Software Engineering*, in press.
10. Mirco Tribastone. The PEPA Plug-in Project. In *Fourth International Conference on the Quantitative Evaluation of Systems*, pages 53–54, Edinburgh, United Kingdom, September 2007. IEEE Computer Society Press.
11. Mirco Tribastone. *Scalable Analysis of Stochastic Process Algebra Models*. PhD thesis, School of Informatics, The University of Edinburgh, 2010.

A Further Remark on Diffusion Approximations with RB and ER Boundary Conditions

Kentaro Hoshi, Yu Nonaka, Yoshitaka Takahashi, and Naohisa Komatsu

Abstract Diffusion approximations are developed different ways, yielding different results. In this paper we redevelop the diffusion approximation for the unfinished work process in the GI/G/1 system with reflecting barrier and elementary return boundary conditions, denoted as DAU(RB) and DAU(ER). The accuracy comparisons are presented among DAU(RB), DAU(ER), and the diffusion approximations for the queue-length process by Heyman, Kobayashi, and Gelenbe; to answer the question which diffusion approximation is the best.

1 Introduction

The diffusion approximations are widely used to evaluate the performance measures in a renewal input, generally distributed service time queueing system, since the Markovian approach sometimes requires an elegant and subtle numerical technique. The queue-length process has been mainly approximated by a diffusion process with either reflecting barrier (RB) or elementary return (ER) boundary condition. The RB boundary condition has been introduced by Heyman [4] and Kobayashi [6], while ER has been considered and generalized by Gelenbe [2]. However, there is almost no literature to answer the question which diffusion approximation will result in the best accuracy, except for Whitt [10], and they are no longer consistent with the ex-

Kentaro Hoshi
Faculty of Science and Engineering, Waseda University, 169-8555 Tokyo, Japan, e-mail: sizer@toki.waseda.jp

Yu Nonaka
Faculty of Science and Engineering, Waseda University, 169-8555 Tokyo, Japan

Yoshitaka Takahashi
Faculty of Commerce, Waseda University, 169-8050 Tokyo, Japan

Naohisa Komatsu
Faculty of Science and Engineering, Waseda University, 169-8555 Tokyo, Japan

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_28,

act result for the Poisson-input single-server (M/G/1) system, as Whitt has already pointed out.In this paper, continuing the effort and spirit of Whitt, we redevelop the diffusion approximations for the unfinished work process in the GI/G/1 system. Our approach is based on a diffusion process for the unfinished-work. Unfortunately, the mean unfinished work is not practical performance measure from a customer's view point. We consider the relationship between the mean unfinished work and the mean waiting time which compensates the pioneering work by Gelenbe [2]. The diffusion approximation for the unfinished work with RB or ER boundary condition will be denoted as DAU(RB) or DAU(ER), respectively. We present the accuracy comparisons among DAU(RB), DAU(ER), and the diffusion approximations by Heyman [4], Kobayashi [6], and Gelenbe [2]; to try to answer the question which diffusion approximation is the best.

2 The GI/G/1 System

There is a single server, unlimited waiting space. We assume the first-in first-out discipline.
A) The customers arrival process forms a renewal process with independent and identically distribute (i.i.d.) inter-arrival time, (A), with arrival rate λ and the squared coefficient of variation as,

$$\lambda = \frac{1}{E(A)} \qquad C_A^2 = \frac{Var(A)}{E(A)^2} = \lambda^2 Var(A)$$

B) The service time (B) is i.i.d. with mean and the squared coefficient variation as,

$$E(B) = \frac{1}{\mu} \qquad C_B^2 = \frac{Var(B)}{E(B)^2} = \mu^2 Var(B)$$

The cumulative service time distribution, $B(t)$, is defined as $B(t) = P(B \leq t)$. The traffic intensity (ρ) is them given by $\rho = \lambda/\mu$. Which is assumed to be less than unity $(\rho < 1)$ for queueing system stability.

3 Diffusion Approximations for the Unfinished Work Process

Let $V(t)$ be the unfinished work process and $f(x,t)$ be the probability density function (*pdf*) of $V(t)$, i.e., $f(x,t)dx \equiv P(x \leq V(t) < x+dx)$.
When we approximate the unfinished work process, we require a boundary condition. The reflecting barrier (RB) and elementary return (ER) boundaries are widely used. Here, we consider both boundary conditions as follows,

A) Diffusion equation with RB [4], [6], [7], [10]
If we set the reflecting barrier (RB) at space origin $(x=0)$, the pdf $f(x,t)$ is assumed to be satisfied as

$$\frac{\partial f}{\partial t} = -\alpha + \frac{\partial f}{\partial x} + \frac{\beta}{2}\frac{\partial^2 f}{\partial x^2} \qquad 0 = \left[-\alpha f + \frac{\beta}{2}\frac{\partial f}{\partial x}\right]_{x=0} \qquad \lim_{x\to\infty} f(x,t) = 0$$

B) Diffusion equation with ER [2], [8]
If we set the elementary return (ER) boundary at space origin $(x=0)$, the pdf $f(x,t)$is assumed to be satisfied as

$$\frac{\partial f}{\partial t} = -\alpha\frac{\partial f}{\partial x} + \frac{\beta}{2}\frac{\partial^2 f}{\partial x^2} + \lambda\pi_0(t)\frac{dB(x)}{dx} \qquad \frac{d\pi_0(t)}{dt} = -\lambda\pi_0(t) + \left[-\alpha f + \frac{\beta}{2}\frac{\partial f}{\partial x}\right]_{x=0}$$

$$\lim_{x\to 0} f(x,t) = \lim_{x\to\infty} f(x,t) = 0$$

Here, the parameters (α,β) are called as the infinitesimal mean and variance, and obtained by Gelenbe [2] as

$$\alpha = \frac{\lambda}{\mu} - 1 = \rho - 1 \qquad \beta = \frac{\lambda}{\mu^2}\left(C_A^2 + C_B^2\right)$$

The steady-state solution pdf $f(x)$ of the diffusion equations are obtained by Heyman [4] under RB, by Gelenbe [2] and corrected by Takahashi [8] under ER as
(RB solution)

$$f(x) = -\frac{2\alpha}{\beta}e^{\frac{2\alpha}{\beta}x} \tag{1}$$

from which we have the mean unfinished work

$$E(V) = \int_0^\infty xf(x)dx = -\frac{\beta}{2\alpha} \tag{2}$$

(ER solution)

$$f(x) = \frac{2\pi_0}{\beta}\lambda\int_0^x (1-B(y))e^{-\frac{2\alpha}{\beta}y}dy\cdot e^{\frac{2\alpha}{\beta}x} \qquad \pi_0 = \alpha \tag{3}$$

from which we have the mean unfinished work as

$$E(V) = \int_0^\infty xf(x)dx = \frac{1}{2}\left[\lambda E(B^2) - \rho\frac{\beta}{\alpha}\right] \tag{4}$$

The mean unfinished work $E(V)$ is the time average, while the mean waiting time $E(W)$ is the customer average. Thus, the Brumelle's law [1] is applied to find the relationship between $E(V)$ and the mean waiting time $E(W)$ as

$$E(W) = \frac{2E(V) - \lambda E(B^2)}{2\rho} \tag{5}$$

It should be noted that Gelenbe [2] has not used this relationship but regarded $E(W)$ as $E(V)$. Substituting the diffusion parameters (α, β) into equations (2) and (4), we have the following explicit formulas from equation (5).
DAU(RB): $E(W)$ under the RB boundary condition is the given by

$$E(W) = \frac{(C_A^2 - 1) + \rho(C_B^2 + 1)}{2\mu(1-\rho)} \tag{6}$$

DAU(ER): $E(W)$ under the ER boundary condition is obtained as

$$E(W) = \frac{\rho(C_A^2 + C_B^2)}{2\mu(1-\rho)} \tag{7}$$

4 Three Other Diffusion Approximations

Before discussing the DAU(RB) and DAU(ER) accuracy, we introduce the major three diffusion approximates for the queue-length process with appropriate boundary conditions. Heyman [4] has presented the diffusion approximation for the queue-length process with reflecting barrier (RB) boundary and obtained the waiting time as

$$E(W) = \frac{C_A^2 + \rho^{-1}C_B^2 - 2(1-\rho)}{2\mu(1-\rho)} \tag{8}$$

Kobayashi [6] also presented the diffusion approximation for the queue-length process with reflecting barrier (RB) boundary, but he has not set at the space-origin (x=0), but (x=1) and the probability that the system's idle is set to be (1-ρ) so that,

$$E(W) = \frac{\hat{\rho}}{\mu(1-\hat{\rho})} \tag{9}$$

where

$$\hat{\rho} = exp\left[\frac{-2(1-\rho)}{(\rho C_A^2 + C_B^2)}\right]$$

Gelenbe [2] has presented the diffusion approximation for the queue-length process with instantaneous return (IR) boundary. IR is an extension of ER in the sense that the sojourn time is generally distributed, but finally, the sojourn time is proved to be one parameter (only mean dependent) distribution, and so, ER where the sojourn time distribution is assumed to be exponential is enough. Anyway Gelenbe has obtained the following waiting time formula.

$$E(W) = \frac{\rho(C_A^2 + 1) + (C_B^2 - 1)}{2\mu(1-\rho)} \tag{10}$$

See also Gelenbe and Mitrani [3] for their heuristic refinement of equation (10).

5 The Accuracy Comparisons

A trivial consistency check mentioned in Whitt [10] is non-negatively. $E(W)$ should be non-negative for all parameter values. Note that Heyman, DAU(RB), Gelenbe can be negative for small C_A^2 and C_B^2.
If we assume the M/G/1 system ($C_A^2 = 1$), it is also natural that diffusion approximations agree with the Pollaczek-Khinchin (P-K) formula;

$$E(W) = \frac{\rho(1+C_B^2)}{2\mu(1-\rho)} \tag{11}$$

However only DAU(RB) and DAU(ER) satisfy this consistency. The queue-length based diffusion approximations have not this consistency. In Gelenbe and Mitrani [3], they have proposed heuristically a refinement such that this consistency is satisfied.
On the other hand, if the service time is exponential distributed or geometrically distributed ($C_B^2 = 1$), the both equations (8) and (12), are reduced to

$$E(W) = \frac{\rho(C_A^2+1)}{2\mu(1-\rho)} \tag{12}$$

which means DAU(ER) and Gelenbe's queue-length based diffusion approximation with IR (ER) boundary have the identical values for the GI/M/1 (or the GI/Geometric/1) system.
As we see the characteristics of DAU (RB) and DAU (ER) in Hoshi et al. [5], we propose the following interpolation between the waiting times by DAU(RB) and DAU(ER)

For $C_A^2 + C_B^2 \geq 1/2, \quad C_A \leq 1,$

$$Proposed = \rho DAU(RB) + (1-\rho)DAU(ER) \tag{13}$$

For $C_A^2 + C_B^2 \geq 1/2, \quad C_A > 1,$

$$Proposed = \frac{\rho(1-C_B)^2}{C_A^2+C_B^2} DAU(RB) + \left\{1 - \frac{\rho(1-C_B)^2}{C_A^2+C_B^2}\right\} DAU(ER) \tag{14}$$

For $C_A^2 + C_B^2 < 1/2,$

$$Proposed = \rho C_A^2 DAU(RB) + (1-\rho C_A^2)DAU(ER) \tag{15}$$

Table 1 shows the mean waiting time, $E(W)$, in the $H_2/E_2/1$ system with exact results by Tijms [9]. For other numerical/simulation comparison results in various queueing systems, see Hoshi et al. [5].

Table 1 The mean waiting time, $E(W)$, in the $H_2/E_2/1$ system. $C_A^2 = 2.0$, $C_B^2 = 0.5$

ρ	Exact	DAU(ER)	DAU(RB)	Proposed	Gelenbe	Kobayashi	Heyman
0.2	0.387	0.313	0.813	0.317	0.063	0.203	1.813
0.5	1.445	1.250	1.750	1.276	1.000	1.055	2.000
0.8	5.281	5.000	5.500	5.134	4.750	4.766	5.563

6 Conclusion

We have shown the mean waiting time result from the diffusion approximation for DAU(ER), cannot be negatively valued unlike those results from the diffusion approximation for DAU(RB), and from the diffusion approximations for the queue-length process by Heyman and Gelenbe. For the GI/M/1 system, we have shown that the mean waiting time result from DAU(ER) coincides with the result from the diffusion approximation for the queue-length process by Gelenbe. We have also seen through our numerical examples, and shown such an appropriate interpolation between DAU(ER) and DAU(RB) as proposed in this paper can yield the best accuracy among all the diffusion approximations, which may lead to a future study topic for applying the diffusion approximation for a more complicated queueing system.

Acknowledgements This research was supported by "Waseda University Global COE Program 'International Research and Education Center for Ambient SoC' sponsored by MEXT, Japan."

References

1. S.L. Brumelle. On the relation between customer and time averages in queues. *J. Appl. Prob. 8*, pages 508–520, 1971.
2. E. Gelenbe. Probabilistic models of computer systems, part II, diffusion approximations, waiting times and batch arrivals. *Acta Informatica*, 12(4): 285–303, 1979.
3. E. Gelenbe and I. Mitrani. Analysis and synthesis of computer systems. *Academic Press*, London and New York, 1980.
4. D. P. Heyman. A diffusion model approximation for the Gl/G/1 queue in heavy traffic. *Bell System Tech. J. 54*, pages 1637–1646, 1975.
5. K. Hoshi. A further refinement of the diffusion approximations for GI/G/1 system. *IECE Technical Report*, CQ, March 2011.
6. H. Kobayashi. Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions. *JACM*, 21: 316–328, 1974.
7. G.F. Newell. *Applications of Queueing Theory, 2nd e.d.* Chapman and Hall, 1982.
8. Y. Takahashi. Diffusion approximation for the single-server system with batch arrivals of multi-class calls. *IECE Transactions*, J69-A (3): 317–324, 1986.
9. H.C. Tijms. *Stochastic Models, an Algorithmic Approach*. John Wiley and Sons, Chichester, 1994.
10. W. Whitt. Refining diffusion approximations for queues. *Operations Research Letters*, 1: 165–169, November 1982.

Stochastic Petri Nets Sensitivity to Token Scheduling Policies

G. Balbo, M. Beccuti, M. De Pierro, and G. Franceschinis

Abstract Stochastic Petri Nets and their generalizations are powerful formalisms for the specification of stochastic processes. In their original definition they do not provide any specification for the token extraction order applied by a transition to its input places, however this aspect may be relevant if timed transitions are interpreted as servers and tokens as customers, so that the extraction order may be interpreted as a scheduling policy. In this paper we discuss how the token scheduling policies different from the Random Order one which is assumed by default, may be relevant for the computation of average performance indices.

1 Introduction

The original definition of Stochastic Petri Nets (SPN) [9] and of their later extensions (e.g. Generalized Stochastic Petri Nets [1] - GSPNs - and Stochastic Well-Formed Nets [5] - SWNs) do not provide any specification for the policy used to extract the tokens from the input place(s) of a transition upon its firing (queueing policy). This is due to the fact that when firing times have negative exponential distribution and tokens are indistinguishable, the specification of queueing policies become inessential as long as average performance indices are the objective of the analysis: hence tokens are picked from the input places in *Random Order* (RO).

When timed transitions are interpreted as servers, firing delays as activity durations, and tokens as customers that are *moved* from input to output places upon transition firings, response time *distributions* may become important and the need arises of specifying queueing policies for input places (an issue that has been studied in [2]) and of accounting for the relevant results of queueing theory that apply within this context. The impact that service policies have on the behavior of queueing models is well understood and average performance indices such as throughputs, utilizations, mean queue lengths, and mean response times are shown to be insensi-

G. Balbo · M. Beccuti · M. De Pierro
Dipartimento di Informatica, Università di Torino, e-mail: {balbo,beccuti,depierro}@di.unito.it

G. Franceschinis, Dipartimento di Informatica, Università del Piemonte Orientale "A. Avogadro", e-mail: giuliana.franceschinis@mfn.unipmn.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 181
DOI 10.1007/978-3-642-20009-0_29,

tive to queueing policies, as long as they are work-conservative and the service time distribution is negative exponential [6]. Things start to be more complex if general service time distributions are considered in conjunction with queueing policies that exhibit preemptive or sharing features. Indeed, it is well known [6, 7, 3] that the average performance indices computed for an M/G/1 queue with Processor Sharing (PS) or Last-Come-First-Served Preemptive-Resume (LCFS-PR) queueing policies are identical to those of M/M/1 queues with same mean service times while they differ considerably when the queueing policy is First-Come-First-Served (FCFS): in this case the coefficient of variation (CV) of the service time distribution plays an important role. Hence, when generally distributed firing times are considered in SPNs (possibly represented through subnets of exponential transitions, in case of phase type distributions [1]) token queueing policies should not be overlooked.

Things become even more intriguing when we turn our attention towards High Level Stochastic Petri nets, such as SWNs, where tokens are *colored* and are thus distinct. In this case both the extraction criteria from the input place(s) and the service time may depend on the token *color*: this is implemented through arc functions and color dependent rates. A peculiar feature of SWNs is a carefully defined syntax, which allows to automatically exploit behavioral symmetries when performing state space based analysis. In case of completely symmetric SWNs and under some additional restrictions (on color synchronization), it may still happen that queueing policies do not influence the average performance indices, however if the color of tokens representing customers are used to model different behaviors (through *static subclasses*) or to synchronize tokens (as in the fork-join example of Sec. 3), the queueing policy may have an impact.

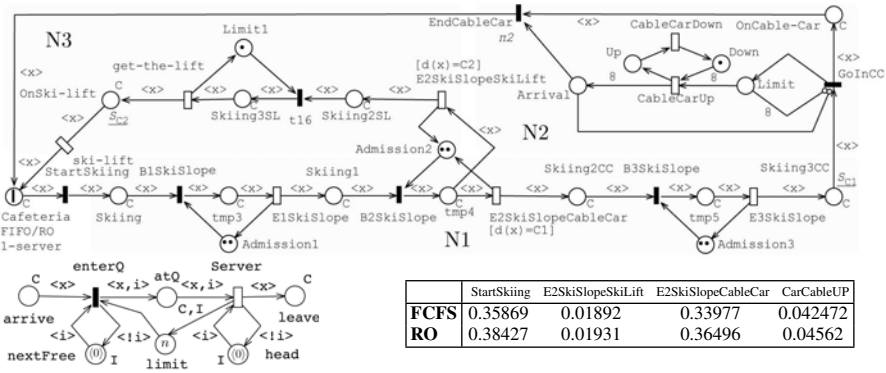
In this paper we address these last issues by presenting the results obtained for a set of SWN models which reflect some of these situations (Sec. 3). In the first example a case is discussed where the queueing policy at a service station with negative exponential service time and color independent rate is relevant, due to batch arrivals of tokens of a subclass. In the second example the effect of a hyper-exponential service time distribution is first considered, then the impact of different queueing policies on the branches of a fork-join structure is studied. Finally we consider a simplified version of a Queueing PN (QPN) model [4] proposed in [8], for which it was observed the importance of the queueing policy of a *buffer place* due to the presence of color dependent service rates.

The examples discussed in this paper are meant to provide evidence that disregarding the order of extraction of tokens from places may produce misleading results. As such, they represent the basis for the characterization of SWN model classes that are insensitive to the queueing policy, on which we are currently working. For the moment we emphasize the importance of adding some *syntactic sugar* to the SWN formalism thus providing some new primitives similar to those proposed in QPNs. In practice, this amounts to the development of compact submodels that can be automatically embedded in the SWN representation of a system: in Sec. 3 we discuss how this can be done without giving up the possibility of using the efficient Symbolic Reachability Graph (SRG) algorithm, that exploits model symmetries and considerably reduces its state space size.

2 Token Queueing Policies: Some Examples

Before discussing the SWN examples, let us introduce some notation used in the models; for more details on the formalism the reader can refer to [5]. The SWN formalism is a flavor of Colored Petri Nets with a well-defined syntax. Tokens in SWN can be distinguished through *colors*, i.e. they may carry some kind of information. In all the models discussed in this paper, we have only one color class: C in the models of Figs. 1 and 2, D in the model of Fig. 3. A class can be partitioned into static subclasses, e.g in Fig. 1, $C = C1 \cup C2$: intuitively colors in the same class have the same nature (e.g. skiers); while those in the same subclass share also the same potential behavior (e.g. skiers using the ski-lift). A color domain is associated with places and transitions. The colors of a place label the tokens that it contains (e.g place *Busy* in Fig. 2 has color domain C). Instead, the color domain of transitions define different *firing modes*. In order to specify the effect of each firing mode, a function is associated with each arc, specifying the colored tokens that are added to, or withdrawn from, the corresponding place when the transition fires in a given mode. A SWN color function is defined combining predefined basic functions: in all the examples only one function is used - $\langle x \rangle$ - which selects an item of a transition color tuple. Finally, transitions can have guards, specified through boolean expressions, whose terms are predefined atomic predicates like $[x = y]$. For example, the guard $[d(x) = C1]$ of transition *E2SkiSlopeCableCar* in Fig. 1 means that this transition may only fire for a color x belonging to static subclass $C1$. All timed transition in our examples will have as many modes enabled in parallel as the number of customers in their input place. In all models we use a notation which is not part of the SWN formalism, but is instead inherited from QPNs: queue places are drawn as circles with a vertical bar and represent a service center for which a queueing policy and a service time distribution are defined. Before proceeding with the solution, these places must be substituted with appropriate subnets, depending on the characteristics of the corresponding service center; subsequently the Continuous Time Markov Chain underlying the model is constructed and analyzed. Fig. 1 (lower) shows an example of subnet corresponding to a queue place with FCFS queueing policy, a single server and negative exponential service time. The queue is modeled by place atQ, with color domain $C \times I$, representing a circular array (I encodes the circularly ordered set of indices). Places *head* and *nextFree* represent the pointers to the first element in queue and to the first empty element, respectively. Transition *Server* has at most one enabled instance at a time, that upon firing removes from place atQ the token $< x, i >$ with i equal to the color of the (unique) token in place *head*, and updates the color of the token in *head* to the successor of i denoted $!i$.

The Impact of Batch Arrivals: The Skiers Model. The customers in this example are skiers that get to the top of a mountain using either a cable-car (subnet $N2$) or a ski-lift (subnet $N3$): hence color class C representing customers, is partitioned in two static subclasses $C1$ and $C2$. When the skiers get to the top of the mountain, they all stop at the cafeteria for a cup of coffee (queue place *Cafeteria*). After drinking the coffee they start skiing down-hill along a common path with sections that, being very narrow, allow only two skiers to proceed "in parallel" (subnet $N1$). Then the



	StartSkiing	E2SkiSlopeSkiLift	E2SkiSlopeCableCar	CarCableUP
FCFS	0.35869	0.01892	0.33977	0.042472
RO	0.38427	0.01931	0.36496	0.04562

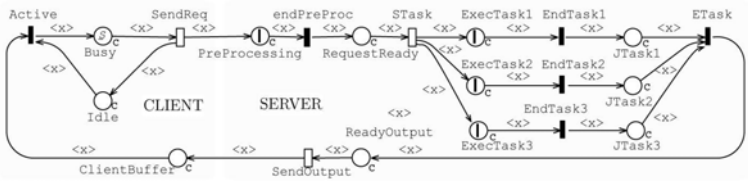
Fig. 1 SWN model for the skiers example (upper fig.); SWN submodel of FCFS policy (lower fig.); Table of the throughputs for RO or FCFS queueing policy at the cafeteria.

ski-lift skiers queue for the lift (place *Skiing2SL*), while the others proceed along the track until they get to the cable-car station (place *Skiing3CC*): the cable-car leaves when K skiers are ready to board.

Our experiments show that *Cafeteria* queueing policy (RO or FCFS) has an effect on the performance indices of the model, despite the service center behavior is completely symmetric w.r.t. the two types of customers. In the table of Fig. 1 we assume $K = 8$, $|C1| = 10$ and $|C2| = 2$ and we compare some throughput obtained with the two queueing policies. These results are due to the fact that the skiers belonging to $C1$ arrive in batch of size 8 to the cafeteria. The difference becomes smaller and smaller as the batch size (K) decreases.

The Effect of Service Time Distribution and of Color-Based Synchronization: The Client-Server Model. The model in Fig. 2 represents a client-server architecture where several similarly behaving clients, modeled by the color class C, can request a remote service from a server. When a client sends a request to the server, the message is queued in the server buffer for pre-processing (queue place *PreProcessing*) and next it is served by generating three threads that are executed in parallel (places $ExecTask_i$). The output is sent back to the client when all three threads end. The first experiment shows that depending on the CV of the pre-processing execution time, the queueing policy of queue place *PreProcessing* may have an impact on the performance indices. In Fig. 2, left table, the system throughput and the mean number of customers in the queueing place *PreProcessing* are compared when $|C| = 6$, the queueing policy of *PreProcessing* is either FCFS or PS, and the pre-processing execution time has a hyper-exponential distribution (rates $\lambda_1 = 1$ and $\lambda_2 = 0.0006622517$ are chosen with probability $p = 0.9998$ and $1 - p$, respectively). The queueing policy of *PreProcessing* is irrelevant when the execution time has negative exponential distribution.

This model allows also to show that the fork/join construct, although completely symmetric, is sensitive to the queueing policies used in its branches. Experiments have been performed with PS queueing policy and negative exponential service time in the queue place *PreProcessing*, as well as queue places *ExecTaski* have either PS



PreProcessing	FCFS	PS
Throughput	0.54889	0.72409
Mean queue length	2.16360	0.71958

	3-PS	2-PS & 1-FCFS	1-PS & 2-FCFS	3-FCFS
System Throughput	0.657712	0.662527	0.669431	0.678263
$\sum_i E[Jtaski]$	3.75654	3.6294	3.4154	3.13833

Fig. 2 SWN model for the Client-Server example; Performance indices of the Client-Server model varying the queueing policy: (left) in PreProcessing with hyper exponential service time; (right) in the fork/join construct.

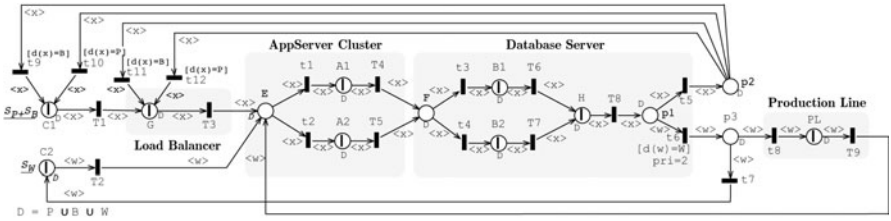
or FCFS queueing policy. In the right table of Fig. 2 we report the system throughput and the sum of the average numbers of tokens in places *JTaski* for these cases.

Distributed Component-Based Model. This example is inspired by the QPN model proposed in [8]; it represents a distributed component-based system consisting of four parts: Load Balancer (LB), Application Server Cluster (APC), Database, and Production Line Station interacting to manage dealer, order entry, and manufacturing applications. The corresponding SWN model is shown in Fig. 3; where the system applications are modeled by color class D which is partitioned in three static subclasses: B dealer applications, P order entry applications and W manufacturing applications. The LB distributes the incoming B and P requests across the nodes in the APC depending on their loads. All service times depend on the type (static subclass) of application. The APC subnet models a cluster which processes the incoming B, P, and W applications. Each node implements a PS queueing policy (places *A1* and *A2*). Processed tasks are submitted to the Database subnet for their execution. Each task execution involves a CPU burst, followed by an I/O burst. The CPU is assigned to each task using a PS policy. Then, a disk access is required for each task according to a FCFS policy (place H). Finally, the last subnet represents the production line station: a w task in place PL models the execution of a manufacturing order by the production line stations.

Similarly to what was observed in [8], the queueing policy applied to the LB component (place G) influences the system performance. In our experiment we studied the effects of the FCFS and PS policies, assuming $|B| = 2$, $|P| = 2$ and $|W| = 3$. The table shown in Fig. 3 reports the throughputs and the average number of customers in some of the queue places.

3 Discussion and Concluding Remarks

The examples presented in the previous section show that when developing SWN models, place queueing policies may have an impact on average performance indices. At first sight this observation may appear obvious, however we believe it is important because the interplay between queueing policies and color-driven behavior, or non-exponential (possibly phase type) firing time distribution, may be overlooked when modelers are used to work with formalisms characterized by undistinguishable tokens and negative exponential firing delays. This suggests that pro-



	Mean queue lengths		Throughputs	
	FCFS	**PS**	**FCFS**	**PS**
Load Balancer	0.501757	0.60422	0.01575875 (67%P 33%B])	0.01482976 (60%P 40%B)
Production Line	1.85466	1.86103	0.00556954 (W)	0.00558867 (W)
Database Disk	0.155609	0.143083	0.02689785 (40%P 19%B 41%W)	0.02600706 (34%P 23%B 43%W)

Fig. 3 SWN model for the distributed component-based example; throughputs and average number of customers computed for different queueing policies at the LB.

viding the modeler with syntactic elements favoring the explicit specification of queueing policies as well as firing time distributions in a compact form, may help to build correct models. Moreover we are interested in embedding these new features in SWN models still preserving the most interesting feature of SWNs, that is their efficient state space analysis algorithms based on symmetries exploitation; the examples considered in this paper were solved by substituting the queue places with relatively simple subnets satisfying the SWN syntax constraints: this transformation could be automated, allowing to exploit the standard efficient analysis algorithms based on the SRG. For instance, in the Client-Server model, the state space reduction achieved by the SRG w.r.t. to the RG is ~ 275 when the queue places $ExecTask_i$ have PS queueing policy; while it is ~ 3000 when their queueing policy is FCFS.

References

1. M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets*. J. Wiley, New York, NY, USA, 1995.
2. G. Balbo, M. Beccuti, M. De Pierro, and G. Franceschinis. First passage time computation in Tagged GSPN with queue places. *Computer Journal*, 2010. Accepted for publication.
3. F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2): 248–260, 1975.
4. F. Bause and P.S. Kritzinger. *Stochastic Petri Nets – An Introduction to the Theory*. Friedr. Vieweg & Sohn Verlag, Braunschweig/Wiesbaden, Germany, Second Edition, 2002.
5. G Chiola, C. Dutheillet, G. Franceschinis, and S. Haddad. Stochastic well-formed coloured nets for symmetric modelling applications. *IEEE Trans. on Computers*, 42(11): 1343–1360, November 1993.
6. E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems*. Academic Press, London, 1980.
7. L. Kleinrock. *Queueing Systems. Volume 1: Theory*. J. Wiley, New York, NY, USA, 1975.
8. Samuel Kounev. Performance Modeling and Evaluation of Distributed Component-Based Systems Using Queueing Petri Nets. *IEEE Trans. Softw. Eng.*, 32(7): 486–502, 2006.
9. M. K. Molloy. Performance analysis using Stochastic Petri Nets. *IEEE Tr. on Computers*, 31(9): 913–917, 1982.

On Lifetime Optimization of Boolean Parallel Systems with Erlang Repair Distributions

Alexander Gouberman and Markus Siegle

We assume a Boolean parallel system with exponentially distributed component failure times. In order to maximize the lifetime of the system we consider a repairman with Erlang-k distributed repair time. By extending the classical exponential case $k = 1$ to $k \geq 2$ different repair semantics arise in this context. In the case of restart repair semantics we show that a repairman should have few Erlang phases in order to maximize mean time to failure (MTTF). In the case $k \geq 2$ with full memory semantics an optimal repairman policy shows counter-intuitive repair behaviour dependent on the mean repair time.

1 Introduction

The field of research for system optimization covers a wide range of models and specialized optimization algorithms. For the subclass of coherent Boolean systems, with which we are concerned here, in [6, 1] the authors maximized the availability of a series system by assigning the repairman to the most reliable failed component (MRC policy). In [2] a generalization of this result to K-out-of-N systems and a repair team consisting of several repairmen was established. The assumption of both exponential component failure rate and repair rate makes it possible to provide an optimal policy which depends only on the order of failure resp. repair rates and not on their concrete values: the fastest repairman should repair the most reliable component. While this assumption may hold for components due to the memoryless property of the exponential distribution (if no concrete component model is known), it is used for the repairman in general just for reasons of model simplification. Repair time distributions which follow the typical "S"-shape like Weibull or Lognormal distributions or even deterministic repair time are known to be more realistic because they often fit to empirical repairman data [3]. Erlang distributions

Alexander Gouberman, Markus Siegle – Fakultät für Informatik, Universität der Bundeswehr München, e-mail: alexander.gouberman@unibw.de, markus.siegle@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 187
DOI 10.1007/978-3-642-20009-0_30,

with a small number of phases can be used to approximate these repair distributions. We extend the exponential repair model to Erlang repair distributions, whereby first of all some repair semantics concerning the Erlang phases have to be chosen. We concentrate on parallel systems and show that for the "restart repair semantics" the fewer phases an Erlang distribution has, the better it maximizes the lifetime of the system. In the case of a "full memory semantics" model we show by a small case study that the optimal policy which assigns the repairman to a failed component depends on the concrete values of the component failure rates and mean Erlang repair time in a somewhat counter-intuitive manner.

2 Markov Decision Model

A parallel system with N components where the lifetime of the i-th component is exponentially distributed with parameter $\mu_i > 0$ can be described by a CTMC with state space $S = \{0,1\}^N$. A system state is given by $x = (x_1, \ldots, x_N)$ where $x_i = 1$ or 0 depending whether the i-th component is functioning or not. Define

$$C_0(x) := \{i \mid x_i = 0\} \quad \text{and} \quad C_1(x) := \{i \mid x_i = 1\}$$

for a state $x \in S$ as the sets of nonfunctioning resp. functioning components. From a state x there are $|C_1(x)|$ transitions given by $x \xrightarrow{\mu_k} (0_k, x),\ k \in C_1(x)$ with rate μ_k, where $(\delta_k, x) := (x_1, \ldots, x_{k-1}, \delta, x_{k+1}, \ldots, x_N),\ \delta \in \{0,1\}$ denotes the state, where the entry corresponding to the k-th component is set to δ. The single absorbing state of the CTMC is $(0, \ldots, 0)$ which represents the failure of the whole parallel system. As a reliability measure for the lifetime of the system we analyze in the following the mean time to failure (MTTF) and assume that the system starts in state $(1, \ldots, 1)$. In order to maximize MTTF by assigning a single repairman with exponential repair time distribution $Exp(\lambda)$ a continuous time Markov decision process (CTMDP) is induced, where the action set in state $x \in S$ is given by $Act(x) := \{r_i \mid i \in C_0(x)\} \cup \{nr\}$, r_i representing the choice to repair the failed component i and nr not to repair any component. The transitions of this CTMDP are given by

$$x \xrightarrow{r_i,\ \lambda} (1_i, x), \text{ for } i \in C_0(x) \quad \text{and} \quad x \xrightarrow{a,\ \mu_j} (0_j, x), \text{ for } j \in C_1(x),\ a \in Act(x),$$

meaning that by choosing the repair action r_i the i-th component can be repaired, and by choosing any action a working component can fail.
The reward which measures the mean sojourn time in a state x when choosing action $a \in Act(x)$ is given by $R(x,a) = \frac{1}{E(x,a)}$, where $E(x,a)$ is the exit rate in state x provided action a is chosen, i.e.

$$E(x,a) = \sum_{j \in C_1(x)} \mu_j + \delta(a)\lambda, \quad \delta(a) \in \{0,1\},\ \delta(a) = 0 \Leftrightarrow a = nr.$$

This reward definition implies that the optimal MTTF from state x to the system failure state $(0, \ldots, 0)$ can be found by solving the Bellman equation induced by this CTMDP. For a detailed discussion on stochastic dynamic programming and especially CTMDPs we refer to [5]. In order to compare exponential and Erlang

repair time distributions with the same expected values we propose two different (extremal) repair model semantics:

- Model 1 ("restart repair semantics"): If during repair of a component a further component fails then the reached phase of the Erlang-k distribution is reset and the repairman can choose any failed component to repair beginning from phase 0 again.
- Model 2 ("full memory semantics"): The Erlang phases of all partially repaired components are remembered and the repairman can be assigned in each state to another failed component (even if no further component failed during the repair procedure). Moreover the repairman continues the repair process from the Erlang phase reached so far for that failed component.

These semantics are indeed "extremal", because one could define other possible repair semantics in between these two. For example, if during a repair procedure of component C a further component fails and the repairman changes to another failed component $\neq C$, then the repair phase of C could get lost. If he does not change he could continue the repair of C from the phase already reached so far.
We now describe the state spaces S_i of model i. In order to compose the system state space S together with an Erlang-k repair state space $E_k := \{0,1,\ldots,k-1\}$ we remember the repair phase for each component, s.t. the composed state space can be described by a subset of $\hat{S} := S \times E_k^N$. For both models 1 and 2 there are states in $\hat{S}$ which are not reachable, more precisely for model 1

$$S_1 = \left\{(x,e) \in \hat{S} \mid x_i = 1 \Rightarrow e_i = 0,\ \exists^{\leq 1} i:\ e_i > 0\right\},$$

meaning that working components cannot be repaired and there is at most one component i which has a positive Erlang repair phase e_i. In the following, we denote a system state by $x \in S$, active resp. failed components by $j \in C_1(x)$ resp. $i \in C_0(x)$ and the Erlang-k parameter by τ. The transitions for model 1 are given by

$$(x,(0,\ldots,0)) \xrightarrow{nr,\mu_j} ((0_j,x),(0,\ldots,0)) \quad \text{and}$$

$$(x,(0,\ldots,e_i,\ldots,0)) \xrightarrow{r_i,\tau} \begin{cases} (x,(0,\ldots,e_i+1,\ldots,0)) & \text{if } e_i < k-1 \\ ((1_i,x),(0,\ldots,0)) & \text{if } e_i = k-1 \end{cases}$$

$$(x,(0,\ldots,e_i,\ldots,0)) \xrightarrow{r_i,\mu_j} ((0_j,x),(0,\ldots,0))$$

In model 2 with full memory semantics there are more reachable states from $\hat{S}$ because Erlang phases are used to model the partially reached repair stage for a failed component. The corresponding state space is given by

$$S_2 = \left\{(x,e) \in \hat{S} \mid x_i = 1 \Rightarrow e_i = 0\right\}$$

with transitions

$$(x,e) \xrightarrow{a,\mu_j} ((0_j,x),e), \text{ for any } a \in Act(x) \quad \text{and}$$

$$(x,e) \xrightarrow{r_i,\tau} \begin{cases} (x,(e_1,\ldots,e_i+1,\ldots,e_N)) & \text{if } e_i < k-1 \\ ((1_i,x),(e_1,\ldots,e_{i-1},0,e_{i+1},\ldots,e_N)) & \text{if } e_i = k-1 \end{cases}$$

Figure 1 shows an excerpt of the CTMDP for a parallel system with $N = 3$ components and Erlang-2 repairman. The upper line represents a system state $x \in S$ and the number below each component state x_i is its Erlang phase $e_i \in E_k$.

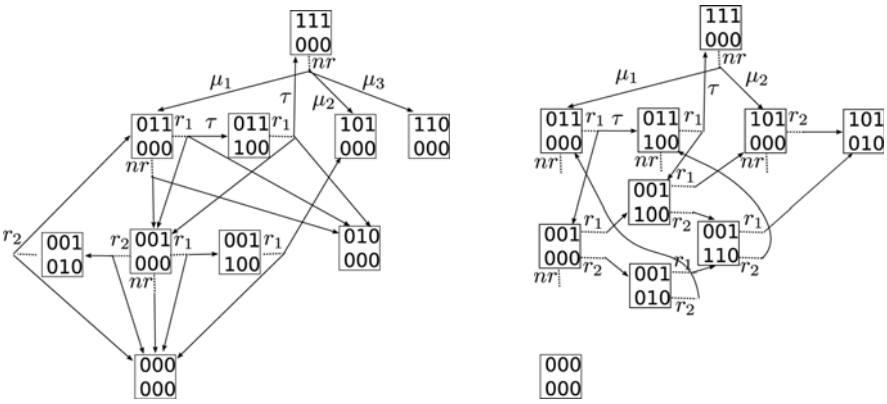


Fig. 1 Excerpt of the CTMDP for different Erlang-2 repair semantics for a parallel system of 3 components ((b) showing especially transitions for repair procedure which differ from (a)). Dashed lines represent decisions and solid lines exponential transitions.

Let us assume, that the repairman does not change to another failed component before a further functioning component fails (as it is always the case in model 1), see Fig. 2. Then the following proposition holds.

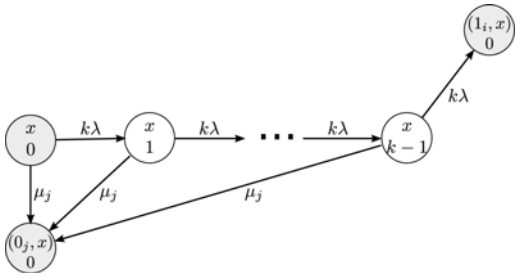


Fig. 2 Local excerpt of the induced CTMC by choosing a repair action r_i in state $x = (0_i, x)$. The lower number indicates the Erlang phase of the i-th component. The shaded states correspond to states in the case of exponential repair time.

Proposition:
Let $x \in S$ and consider Erlang-k distributed repair times $T_k \sim Erl_k(k\lambda)$ with same expected values $\mathbb{E}(T_k) = \frac{1}{\lambda}\ \forall k \in \mathbb{N}$.

(a) The probability to repair a failed component $i \in C_0(x)$ without further failure of another component during the repair procedure is decreasing with k.
(b) The mean time to repair a failed component, under the condition that no further failure of another component during the repair procedure occurs, is increasing with k.

Proof: The probability for following the path

$$\sigma_k := (x,0) \to (x,1) \to \cdots \to (x,k-1) \to ((1_i,x),0)$$

is given by $P_k = \left(\frac{k\lambda}{k\lambda+\mu}\right)^k$, where $\mu = \sum_{j \in C_1(x)} \mu_j$. It is known that $P_k = \left(1+\frac{\mu/\lambda}{k}\right)^{-k}$ is monotonically decreasing (and converging to $e^{-\mu/\lambda}$), thus (a) holds. The mean time to repair a failed component on the path σ_k can be computed by summing up the sojourn times in the states along σ_k: $\mathbb{E}(\sigma_k) = k \cdot \frac{1}{k\lambda+\mu} = \frac{1}{\lambda}\frac{k\lambda}{k\lambda+\mu}$. But since $f(x) := \frac{x}{x+\mu}$, $x \in \mathbb{R}$ is strictly monotonic increasing, statement (b) also holds. □

Special cases: Among all Erlang repair distributions with identical expected values the following holds.

1. An exponentially distributed repair time ($k = 1$) maximizes the probability to repair a failed component and minimizes the mean time to repair it before a further component fails.
2. In case of deterministic repair time (for $k \to \infty$) this probability is minimal $\left(e^{-\mu/\lambda} = P(\text{sojourn time in state } x \text{ is lower than repair time } \frac{1}{\lambda})\right)$ and the corresponding repair time is maximal.

For the restart repair model, exponentially distributed repair time is the best among all Erlang repair time distributions. But since the practical issue of this model is low (and only introduced for comparison with the exponential repairman case) we adhere to the more practical model 2 with full memory semantics, since Erlang phases can approximately describe repair stages during repair of a failed component. For the case $k = 1$ Katehakis and Melolidakis showed, that the MRC policy which assigns the repairman to the failed component with least failure rate maximizes MTTF [2]. We show that for the case $k \geq 2$ in model 2 this is not the case, since repair phases are remembered. Figure 2 shows optimal policies with regard to MTTF maximization in model 2 with $k = 2$ Erlang repair phases for a parallel system with 3 components and component failure rates $\mu_1 = 1,\ \mu_2 = 2,\ \mu_3 = 3$. In this case the optimal policy differs only for state $010 - 001$ in a noncoherent manner: Relatively "slow" and "fast" repairmen should continue to repair the third component, but there is also a region, in which it is better to change repair to component one which is in repair phase 0. The results were computed by implementing the policy iteration method [5] in Wolfram Mathematica 7 and apply it to the CTMDP of model 2.

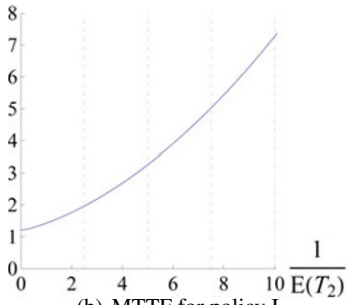
3 Conclusion

We have generalized the classical exponential repairman models to Erlang-k repairman models in order to maximize the MTTF of a Boolean parallel system. In this context we discussed two different semantics which lead to different extensions of the system state space. We showed by optimizing the CTMDP for a toy example with full memory semantics that classical results on optimal repair strategies do not hold any more for Erlang-k repairmen. It would be interesting to see whether avail-

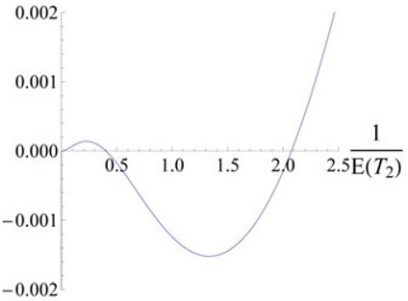
State (x,e)	action set $Act(x,e)$	policy I	policy II
111-000	nr	nr	nr
011-000	nr, r_1	r_1	r_1
011-100	nr, r_1	r_1	r_1
101-000	nr, r_2	r_2	r_2
101-010	nr, r_2	r_2	r_2
110-000	nr, r_3	r_3	r_3
110-001	nr, r_3	r_3	r_3
001-000	nr, r_1, r_2	r_1	r_1
001-100	nr, r_1, r_2	r_1	r_1
001-010	nr, r_1, r_2	r_2	r_2
001-110	nr, r_1, r_2	r_1	r_1
010-000	nr, r_1, r_3	r_1	r_1
010-100	nr, r_1, r_3	r_1	r_1
010-001	$\mathbf{nr}, \mathbf{r_1}, \mathbf{r_3}$	$\mathbf{r_3}$	$\mathbf{r_1}$
010-101	nr, r_1, r_3	r_1	r_1
100-000	nr, r_2, r_3	r_2	r_2
100-010	nr, r_2, r_3	r_2	r_2
100-001	nr, r_2, r_3	r_3	r_3
100-011	nr, r_2, r_3	r_2	r_2

(a) Optimal policies with mean repair time $\mathbb{E}(T_2)$

mean repair time $\mathbb{E}(T_2)$	4	1	0.4	0.1
optimal MTTF	1.244	1.429	1.966	7.261
optimal policy	I	II	I	I



(b) MTTF for policy I



(c) Difference of MTTF values for policies I and II

Fig. 3 Optimal policies for a parallel system with 3 different components and Erlang-2 distributed repair time T_2 with full memory semantics

able measures for component importance (like Barlow-Proschan or Natvig [4]) lead to good heuristics or even optimal repair policies.

References

1. Michael N. Katehakis and Cyrus Derman. Optimal repair allocation in a series system. *Mathematics of Operations Research*, 9(4): 615–623, 1984.
2. Michael N. Katehakis and Costis Melolidakis. Dynamic repair allocation for a k-out-of-n system maintained by distinguishable repairmen. *Probability in the Engineering and Informational Sciences*, 2: 51–62, 1988.
3. R.M. Kieckhafer, M.H. Azadmanesh, and Y. Hui. On the sensitivity of NMR unreliability to non-exponential repair distributions. In *Fifth IEEE International Symposium on High Assurance Systems Engineering*, pages 293–300, 2000.
4. Bent Natvig and Jorund Gasemyr. New results on the Barlow-Proschan and Natvig Measures of Component Importance in Nonrepairable and Repairable Systems. *Methodology and Computing in Applied Probability*, 11: 603–620, 2009.
5. Martin L. Puterman. *Markov Decision Processes*. John Wiley & Sons INC., 1994.
6. Donald R. Smith. Optimal repair of a series system. *Operations Research*, 26(4): 653–662, 1978.

Computing Lifetimes for Battery-Powered Devices

Marijn Jongerden and Boudewijn Haverkort

Abstract The battery lifetime of mobile devices depends on the *usage pattern* of the battery, next to the discharge rate and the battery capacity. Therefore, it is important to include the usage pattern in battery lifetime computations. We do this by combining a stochastic workload, modeled as a continuous-time Markov model, with a well-known battery model. For this combined model, we provide new algorithms to efficiently compute the expected lifetime and the distribution and expected value of the delivered charge.

1 Introduction

The usage of wireless devices like cell phones, laptop computers or wireless sensors is often limited by the lifetime of the included batteries. The lifetime naturally depends on the capacity of the battery and the rate at which it is discharged. However, it also depends on the discharge pattern. When a battery is continuously discharged, a high current will cause it to provide less energy until the end of its lifetime than a low current. This is the so-called *rate-capacity* effect. On the other hand, during periods of low or no current the battery can recover partly. This is the so-called *recovery-effect*. To properly model the impact of the usage pattern on the battery, one has to combine a workload model with a battery model.

We combine the Kinetic Battery Model (KiBaM) [4], which is the simplest model that includes both the above effects, with a continuous-time Markov model to describe the usage pattern of the device, that is, its workload.

Marijn Jongerden
University of Twente, Enschede, The Netherlands e-mail: jongerdenmr@ewi.utwente.nl

Boudewijn Haverkort
Embedded Systems Institute, Eindhoven, The Netherlands e-mail: boudewijn.haverkort@esi.nl
University of Twente, CTIT, Enschede, The Netherlands, e-mail: brh@ewi.utwente.nl

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_31,

In [1] we have proposed this model to compute battery lifetime *distributions*. Here we extend our analysis in order to efficiently compute the distribution of the total charge delivered by the batteries, as well as the expected battery lifetime and the expected charge delivered. The details of this analysis are given in [3].

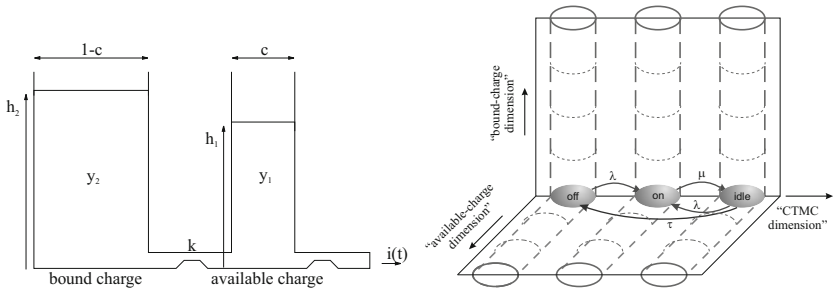
2 Kinetic Battery Model

We use the Kinetic Battery Model (KiBaM) [4] to model the battery. This model is the simplest model that includes the two important non-linear battery properties, the rate-capacity effect and the recovery effect [2].

In the KiBaM the battery charge is distributed over two wells: the available-charge well and the bound-charge well (cf. Figure 1(a)). A fraction c of the total capacity is put in the available charge well, and a fraction $1-c$ in the bound charge well. The available charge well supplies electrons directly to the load ($i(t)$), whereas the bound-charge well supplies electrons only to the available-charge well. The charge flows from the bound charge well to the available charge well through a "valve" with fixed conductance, k. Next to this parameter, the rate at which charge flows between the wells depends on the height difference between the two wells. The heights of the two wells are given by: $h_1 = \frac{y_1}{c}$ and $h_2 = \frac{y_2}{1-c}$. The change of the charge in both wells is given by the following system of differential equations:

$$\begin{cases} \frac{dy_1}{dt} = -i(t) + k(h_2 - h_1), \\ \frac{dy_2}{dt} = -k(h_2 - h_1), \end{cases} \tag{1}$$

with initial conditions $y_1(0) = c \cdot C$ and $y_2(0) = (1-c) \cdot C$, where C is the total battery capacity. The battery is considered empty as soon as there is no charge left in the available charge well, that is, as soon as $y_1 = 0$.



(a) Kinetic Battery Model (b) MRKiBaM

Fig. 1 The two well Kinetic Battery Model and the Markov Reward KiBaM

3 Markov Reward KiBaM

3.1 Introduction

We combine the KiBaM with a stochastic workload, modeled as a continuous-time Markov chain (CTMC), creating a Markov reward model (MRM). Each state in the CTMC represents a mode in which the device can be used, with its own discharge current. Mode switches are represented by the transitions between the states in the Markov chain. In combining the CTMC with the KiBaM, the differential equations of the KiBaM are integrated into the Markov model as accumulated rewards, which represent the levels of charge in the two charge wells. A schematic picture of the combined model is given in Figure 1(b). The first accumulated reward $Y_1(t)$ represents the available-charge well, the second accumulated reward $Y_2(t)$ represents the bound-charge well. The corresponding rates are derived from the KiBaM differential equations (1). Let I_i be the energy consumption rate in a state $i \in S$. The first reward rate then is

$$r_{i,1}(y_1, y_2) = \begin{cases} -I_i + k \cdot (\frac{y_2}{1-c} - \frac{y_1}{c}), & \frac{y_2}{1-c} > \frac{y_1}{c} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

and the second reward rate is

$$r_{i,2}(y_1, y_2) = \begin{cases} -k \cdot (\frac{y_2}{1-c} - \frac{y_1}{c}), & \frac{y_2}{1-c} > \frac{y_1}{c} > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

The interesting question for battery-powered devices is: "When does the battery get empty?" For the Markov Reward KiBaM model, the battery is empty at time t if the available-charge well is empty, that is, if the accumulated reward $Y_1(t) = 0$. Since the accumulated reward $Y_1(t)$ is a random variable, we can only indicate the *probability* that the battery is empty at time t:

$$\mathbb{P}\{\text{battery empty at time } t\} = \mathbb{P}\{Y_1(t) = 0\} \tag{4}$$

The *lifetime* L of a battery is the instant the battery gets empty for the first time, $L = \min\{t \mid Y_1(t) = 0\}$.

3.2 Battery Lifetime

In [1] we showed that one can approximate the MRM with a CTMC by applying a discretization to the accumulated reward levels. In fact, we approximate the continuous accumulated reward growth with a discrete stochastic equivalent in which the accumulated reward is regarded as a discrete Erlang-distributed random variable. This CTMC approximation provides in fact a phase-type distribution for a given workload model. Its absorbing states are the ones where the battery is per-

ceived empty, that is, where the available charge Y_1 reaches zero. Such state is of the form $(i,0,j_2)$, where i is an original MRM state, $j_1 = 0$ represents the empty available-charge well, and j_2 is the discretized level of the charge remaining in the bound-charge well.

The generator matrix of this new CTMC $\mathbf{Q}^*$ can be arranged in such a way that

$$\mathbf{Q}^* = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{T^0} & \mathbf{T} \end{pmatrix}, \tag{5}$$

where $\mathbf{T}$ contains the rates of transitions between non-absorbing states, and $\mathbf{T^0}$ is the matrix with the rates from each non-absorbing state to the absorbing states, which indicate that the battery is empty. If we merge all absorbing states into one, the generator matrix reduces to:

$$\mathbf{Q}^* = \begin{pmatrix} 0 & 0 \\ \underline{T}^0 & \mathbf{T} \end{pmatrix}, \tag{6}$$

where $\underline{T}^0$ is a column vector with the cumulative rates to the absorbing states. The represented phase-type distribution is the approximate distribution for the random variable describing the time it takes for the battery to be emptied.

The expected value of a random variable L having a phase-type distribution is described by $\mathbb{E}[L] = -\underline{\alpha}\mathbf{T}^{-1}\underline{1}$, where α is the initial distribution, and $\underline{1}$ is a column vector of appropriate size with each element equal to one. Thus, if we solve the system of linear equations $\underline{x}\mathbf{T} = -\underline{\alpha}$, we have that $\mathbb{E}[L] = \sum_i x_i$. Using this approach we can approximate the expected battery lifetime for a given workload.

3.3 Delivered Charge

The amount of charge that is actually delivered by the battery depends on the workload. When we look at the absorbing states of the approximating CTMC, we see that if the CTMC ends up in a state $s_a = (i,0,j_2)$, it means that the delivered charge is approximately $C - j_2\Delta$, where Δ is step size of the discretization that is applied to the charge. We can thus compute approximations to the distribution and expected value of the delivered charge.

Since for these computations the time until absorption is not important, it suffices to consider the embedded discrete-time Markov chain with probability matrix $\mathbf{P}^*$, where

$$P^*_{s,s'} = \begin{cases} 1, & \text{if } s = s' \text{ and } Q^*_{s,s} = 0, \\ \frac{Q^*_{s,s'}}{-Q^*_{s,s}}, & \text{if } s \neq s' \text{ and } Q^*_{s,s} \neq 0, \\ 0, & \text{elsewhere.} \end{cases} \tag{7}$$

Following the notation introduced for phase-type distributions we can arrange $\mathbf{P}^*$ such that

$$\mathbf{P}^* = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R^0} & \mathbf{R} \end{pmatrix}. \tag{8}$$

The probability A_{s,s_a} to end in a certain absorbing state s_a, having started in state s is determined by the following system of linear equations:

$$A_{s,s_a} = \begin{cases} 1, & \text{if } s = s_a, \\ \sum_z P^*_{s,z} A_{z,s_a}, & \text{otherwise.} \end{cases} \tag{9}$$

If $\mathbf{B}$ is the matrix consisting of the values A_{s,s_a} where s is a transient state, this system of linear equations can be written as:

$$\mathbf{RB} + \mathbf{R^0 I} = \mathbf{B} \quad \text{or} \quad (\mathbf{I} - \mathbf{R})\,\mathbf{B} = \mathbf{R^0}. \tag{10}$$

This system can be solved for one column of $\mathbf{R^0}$ at a time using standard solution algorithms [5]. The complete matrix $\mathbf{A}$ is obtained by extending the computed matrix $\mathbf{B}$ to include also the absorbing states as initial states:

$$\mathbf{A} = \begin{pmatrix} \mathbf{I} \\ \mathbf{B} \end{pmatrix}. \tag{11}$$

Multiplying the initial distribution vector $\underline{\alpha}$ with $\mathbf{A}$ gives the probability distribution $\underline{a}$ to end up in the different absorbing states, *i.e.*, $\underline{a} = \underline{\alpha}\mathbf{A}$. The element $a_{(i,0,j_2)}$ denotes the probability that the battery gets empty with a residual charge of $j_2\Delta$, and thus having delivered a charge of $C - j_2\Delta$. In doing so, we obtain the distribution of the delivered charge. The expected delivered charge $\mathbb{E}[dC]$ is given by:

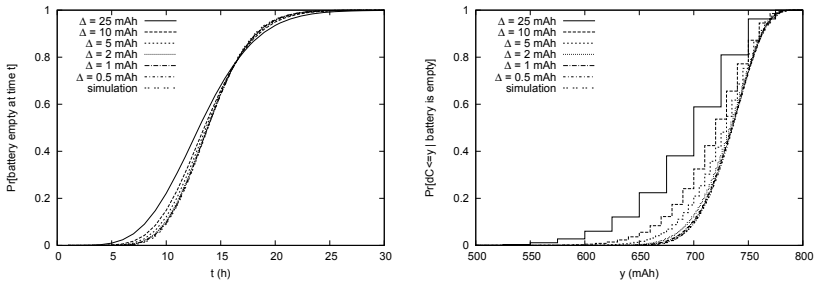
$$\mathbb{E}[dC] = C - \sum_{\substack{(i,0,j_2) \\ \text{is absorbing}}} j_2 \Delta a_{(i,0,j_2)}. \tag{12}$$

4 Results

For the results we use the simple workload model and battery that were also used in [1]. The simple workload model consists of three states, as depicted in Figure 1(b). Initially, the model is in `idle` state. With rate $\lambda = 2$ per hour the model moves into the `on` state. On average, the model spends 10 minutes in the `on` state, resulting in a outgoing rate of $\mu = 6$ per hour. From the `idle` state the device can also move into a power-saving `off` state, this is done – on average – once per hour ($\tau = 1$). The power-consumption rate is low when idling ($I_0 = 8$ mA), it is high when sending data ($I_1 = 200$ mA) and negligible in the sleep state ($I_2 = 0$ mA).

In Figure 2, distributions of the battery lifetime and delivered charge are given for various values of the discretization parameter Δ, and compared to the distributions obtained by simulation. The simulation distributions are based on 10000 runs.

In Figure 2(a) we see that the lifetime distribution is well approached already with $\Delta = 10$ mAh; for the delivered charge distribution Δ has to be decreased to 2 mAh to obtain a good approximation, cf. Figure 2(b).



(a) lifetime distribution (b) delivered charge distribution

Fig. 2 Battery lifetime and delivered charge distribution for a simple workload model

Next to the distributions also the expected values of the battery lifetime and the delivered charge have been computed. For both the expected lifetime and the expected delivered charge the results are similar. The expected values for $\Delta = 0.5$ mAh are 1% lower than the averages obtained from the simulation.

More results can be found in [3], where the simple model is compared with a more complex burst model, and the difference in lifetime and delivered charge between the two models is discussed.

5 Conclusion

We presented a new approach to compute distributions and expected values of both battery lifetime and delivered charge for random workloads. This is an extension to the work presented in [1], where only the lifetime distribution was computed. The results for a simple workload model show that the approach leads to a good approximation for both distributions and expected values.

References

1. L. Cloth, B. R. Haverkort, and M. R Jongerden. Computing battery lifetime distributions. In *Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN '07)*, pages 780–789. IEEE Computer Society Press, 2007.
2. M. R. Jongerden and B. R. Haverkort. Which battery model to use? *IET Software*, 3(6): 445–457, December 2010.
3. Marijn Jongerden. *Model-Based energy analysis of battery powered systems*. PhD thesis, University of Twente, expected dec 2010.
4. J.F. Manwell and J.G. McGowan. Lead acid battery storage model for hybrid energy systems. *Solar Energy*, 50: 399–405, 1993.
5. William J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

I.7 Business Informatics and Artificial Intelligence

Chair: Prof. Dr. Andreas Fink (Helmut Schmidt Universität Hamburg)

Decision support systems are computer-based information systems that are designed with the purpose of improving the process and outcome of decision making. This includes providing data to represent, describe and evaluate the decision problem as well as algorithms to aid in solving these problems. Depending on the kind of decision situation (e.g. with respect to structured vs. semi-structured problems with consequences for the degree of automation that one aims at) different approaches are suitable. This includes data warehousing and business intelligence, interactive decision aids, (meta-)heuristic search procedures, decentralized and agent-based coordination mechanisms, as well as further methods from the field of artificial intelligence.

Furthermore, there is the issue of supporting the practical implementation of such methods. This includes techniques for integrating quantitative methods within business application systems and coupling elements of decision support systems (such as data storage, solver, and user interface) within distributed systems such as the Internet/Web.

We invite the submission of original research contributions, theoretical papers as well as practical case studies, from this spectrum of topics.

A Genetic Algorithm for Optimization of a Relational Knapsack Problem with Respect to a Description Logic Knowledge Base

Thomas Fischer and Johannes Ruhland

Abstract We present an approach that integrates a description logic based knowledge representation system into the optimization process. A description logic defines concepts, roles (properties) and object instances for relational data, which enables one to reason about complex objects and their relations. We outline a relational knapsack problem, which utilizes the knowledge base during optimization. Furthermore, we present a genetic algorithm to outline an approximate algorithm for a heuristic solution.

1 Problem Description and Formulation

In this paper we present an approach that integrates a knowledge representation system into the optimization process. A knowledge-based system is able to find implicit consequences of its explicitly represented knowledge. The idea is to make use of the explicit and implicit part of the knowledge representation to guide the optimization, which distinguishes it from traditional approaches that access relational databases and flat files. Furthermore, our intention is to achieve a separation of concerns to master complexity. This means that knowledge representation and reasoning are separated from the mathematical problem description as much as possible. In addition, our approach provides an important methodology for the integration of optimization with the semantic computing landscape, which is an important part of the future (distributed) computing in enterprises and the Web [11].

We outline our approach by a knapsack problem (KP). However, our general methodology is not limited to this. In our KP we have different items with a value (or utility) and weight that belong to a class (e.g. *Book*, *ElectronicDevice*). Further-

Thomas Fischer and Johannes Ruhland
School of Economics and Business Administration, Friedrich-Schiller-University Jena,
Carl-Zeiss-Str.3, 07743 Jena,
e-mail: fischer.thomas@uni-jena.de, johannes.ruhland@uni-jena.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_32,

more, there can be subtypes of classes. We assume that the knapsack value is not a simple sum of item values, because objects in the knapsack can build a collection (e.g. collection of artwork), which increases the value[1]. The goal is to find a subset of the n items such that the total value is maximised and all resource constraints are satisfied. In general there can be several relations between objects that influence the knapsack value, which distinguishes it somewhat from a standard quadratic knapsack problem [9, 6], thus we call this a relational knapsack problem (RKP).

Knapsack problems are widely studied due to their ability to closely represent real world problems [6]. In this context, genetic algorithms are often utilized, because they provide the ability of parallel processing, global search and local search as well as rapidity and robustness even for complex problems [15]. Such GAs has been outlined for the multi-dimensional knapsack problem [3, 2, 4] as well as quadratic knapsack problem [10]. However, their domain specific knowledge is fully directly implemented in the GA. In contrast, our approach integrates explicit as well implicit knowledge (logical reasoning) through an external knowledge representations system into the optimization process to master complexity.

Knowledge Representation

The knowledge representation of our approach is based on the logical formalism of description logic (DL). DLs offer a high level of expressivity while still being decidable. A typical DL knowledge base is structured by a so called $TBox$ and $ABox$ ($KB = (TBox, ABox)$. The $TBox$ contains intensional knowledge (terminology of the vocabulary) and is build through the definition of concepts and properties (also called roles). The relationships among the concepts form a lattice like structure. The $ABox$ contains extensional knowledge and denotes knowledge specific to the individuals (assertions). Knowledge representation systems, which are structured by $ABox$ and $TBox$ go beyond the specification of concepts, roles and assertions. They provide reasoning services, which are able to extract implicit knowledge. The interested reader should refer to Baader et al. [1] for a detailed investigation of DLs.

The $TBox$ of our RKP specifies a concept $Object$ as well as various sub-concepts such as $Book$, $ElectronicDevice$ etc. Furthermore, each object has a $hasValue$ and $hasWeight$ property. A special relation is the symmetric property inC ($inC \equiv inC^{-}$), which defines that two objects build a collection. For a reasoner this means that if an object $o1$ is in a collection with $o2$ the reverse is implicit true too. We also specified a concept $Knapsack$ and a relation inK, which defines that an object can be in a knapsack. The $ABox$ defines the actual instances of the $TBox$. We assume that we have different objects, e.g the books named DL or $Software$, as well as their values and weights.

[1] In a project selection problem it would be possible to specify types of projects as well as an increasing overall return, if two or more projects have synergistic effects.

TBox

$Object \sqsubseteq \top$	$\geqslant 1\, hasValue \sqsubseteq Object$
$Knapsack \sqsubseteq \top$	$\geqslant 1\, hasWeight \sqsubseteq Object$
$Book \sqsubseteq Object$	$\geqslant 1\, hasCapacity \sqsubseteq Knapsack$
$ElectronicDevice \sqsubseteq Object$	$\geqslant 1\, inK \sqsubseteq Object$
$CellPhone \sqsubseteq ElectronicDevice$	$\geqslant 0\, inC \sqsubseteq Object$
$Camera \sqsubseteq ElectronicDevice$	$\top \sqsubseteq \forall hasValue.Double$
$Photo \sqsubseteq ElectronicDevice$	$\top \sqsubseteq \forall hasWeight.Double$
$MP3Player \sqsubseteq ElectronicDevice$	$\top \sqsubseteq \forall hasCapacity.Double$
$MusicalInstrument \sqsubseteq Object$	$\top \sqsubseteq \forall inK.Knapsack$
$Computer \sqsubseteq ElectronicDevice$	$\top \sqsubseteq \forall inC.Object$
$Notebook \sqsubseteq ElectronicDevice$	$\top \sqsubseteq \leqslant 1\, hasValue$
$Jewellery \sqsubseteq Object$	$\top \sqsubseteq \leqslant 1\, hasWeight$
$Necklace \sqsubseteq Jewellery$	$\top \sqsubseteq \leqslant 1\, hasCapacity$
$Earring \sqsubseteq Jewellery$	$\top \sqsubseteq \leqslant 1\, inK$
$Bracelet \sqsubseteq Jewellery$	$inC \equiv inC^{-}$
$Charm \sqsubseteq Jewellery$	

ABox

$Knapsack(K1)$	$hasCapacity(K1,720)$	
$Book(DL)$	$hasValue(DL,70)$	$hasWeight(DL,200)$
$Book(Software)$	$hasValue(Software,80)$	$hasWeight(Software,150)$
$Earring(WDiamond)$	$hasValue(WDiamond,50)$	$hasWeight(WDiamond,50)$
$Bracelet(TitanA)$	$hasValue(TitanA,10)$	$hasWeight(TitanA,20)$
$Charm(ItalianH)$	$hasValue(ItalianH,20)$	$hasWeight(ItalianH,50)$
$Ring(KDiamond)$	$hasValue(KDiamond,10)$	$hasWeight(KDiamond,50)$
$Notebook(EEPC)$	$hasValue(EEPC,1000)$	$hasWeight(EEPC,500)$
$Photo(PhotoA)$	$hasValue(PhotoA,150)$	$hasWeight(PhotoA,200)$
	$inC(WDiamond,KDiamond)$	$inC(TitanA,ItalianH)$

We achieve direct access to the knowledge base during optimization through the specification of logical conjunctive queries, which are of the general form

$$(x_1,x_2,\ldots,x_k).\exists(x_{k+1},\ldots,x_m).A_1,\ldots,A_r. \tag{1}$$

$x_1,x_2,\ldots,x_k$ are result variables, whereas $x_{k+1},\ldots,x_m$ will be bound. $A_1,\ldots,A_r$ are atomic formulae that can be combined through conjunction and existential quantification. The query $(v_{o_1}).\exists(o_1).hasValue(o_1,v_{o_1})$ returns a list of values v_{o_1} for any existing object o_1. The result on the previous KB would be $70,80,50,\ldots$. The query $\exists(o_1,o_2,k_1).inC(o_1,o_2) \wedge inK(o_1,k_1)$ is a boolean query and returns 1 for *true* and 0 for $false$. The result on the previous KB would be 0, because currently there is no object in a knapsack.

There are some expression that cannot be stated in a DL. Therefore rules can be added, which make the queries in general undecidable. However, some type of rules do not limit decidability and can be efficiently processed by reasoners [8]. The following rule enables us to specify that every item of the class *Jewellery* builds a collection with any other jewellery item.

$$Jewellery(j),Jewellery(z) \rightarrow inC(j,z) \tag{2}$$

Note, in the *ABox* we created statements for *inC* for specific objects. In addition, we now have stated a more general rule. Both approaches will be adhered in the logical

reasoning. The advantage is that we can simply drop the rule in our KB, if the rules is not longer valid to our problem. The description logic has been implemented in the semantic web standard OWL2 [14] and encoded in the RDF [13] syntax. The expressiveness of our knapsack DL is $\mathcal{ALHIF}(\mathcal{D})$ (see [1]). The Rule (Eq. 2) has been encoded in SWRL [5].

Mathematical Program

With the given framework it is possible to define the mathematical program[2] for the relational knapsack problem.

$$\max_{o_1,o_2 \in Object} \sum_{o_1} (v_{o_1}).hasValue(o_1, v_{o_1}) \times inK(o_1, "K1") + \sum_{o_1} \sum_{o_2} 30 \times (inC(o_1, o_2) \wedge inK(o_1, "K1") \wedge inK(o_2, "K1")) \quad (3)$$

subject to

$$\sum_{o \in Object} (w_o).hasWeight(o, w_o) \times inK(o, "K1") \leqslant (c).hasCapacity("K1", c) \quad (4)$$

and

$$\mathbf{KB} = (\mathbf{TBox}, \mathbf{ABox}) \quad (5)$$

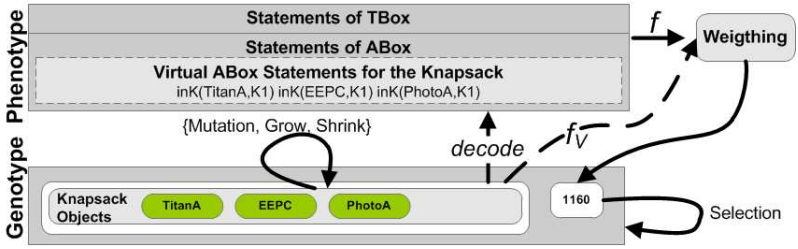
We have assumed that for a collection of two objects the knapsack value increases by 30. It would be easy to extend this program to more knapsacks by adding new *ABox* instances and by adding new resource properties for the objects, which can be also specific to classes of objects, resulting in a relational multiple knapsack problem. The queries are evaluated directly on the knowledge base at runtime to ensure usage of the most current information and to avoid data transformation, if the *ABox* is part of constant change. The query based approach allows utilization of reasoning in the KB and reduces complexity, because the domain knowledge is separated from the mathematical problem.

2 Genetic Algorithm

For a proof of concept, a genetic algorithm (GA) seems to be a suitable solution strategy, because it needs some kind of domain knowledge to avoid a blind search. The overall objective is to search for a set of *inK* statements that optimize the given objective function f in accordance with the KB. The genotype representation is a

[2] There are no existential quantifications in the queries, because the variables are bound to a specific value (e.g. due to the sum operator).

list of objects that should be in the knapsack. The algorithm starts with a random set of possible candidate knapsacks. Promising ones are selected for a pairwise recombination through an implemented standard tournament selection and mutated afterwards. The final number of items in the knapsack is not known in advance, therefore growing and shrinking is important. The new solution candidates are then weighted and added to the population. A standard truncation selection determines the survival of the fittest for the next generation of the population. The weight of an individual is determined by the function f_V, which is an implemented version of the objective function (Eq. 3). In addition we added a penalty for individuals that break the resource constraint (Eq. 4). We use the query language SPARQL to perform conjunctive queries. The implementation of our GA is based on Java and the Jena [7] semantic web library. Furthermore, we use Pellet [12] as the logical reasoner that is also capable of utilizing SWRL [5] rules.



We have evaluated our approach based on three scenarios. The first case is the standard knapsack problem without any instance of the *inC* relation. The other two problems are RKP with or without the logical rule (Eq. 2). We tested our GA random 10 times for each case[3]. The optimum has been computed with CPLEX through materializing all statements and rules of the KB and transforming it to a standard program with linear or quadratic function. However, the transformation becomes increasingly complex, if there would be more logical relations as well as implicit information that have to be adhered in the optimization.

	10 random GA runs					CPLEX	
Case	best	times best	incorrect	mean ± std.dev	skew	Solution	Type
Standard KP	1160	8	1	1079 ±253	−3,2	1160	linear
RKP without Rule	1210	3	2	1026 ±335	−1,9	1210	quadratic
RKP with Rule	1570	3	0	1375 ±148	+0,34	1570	quadratic

As expected the performance of the GA slows down, if the problem becomes harder (inherently non-linear). Some incorrect solutions with penalty affected mean and standard deviation negatively, but more importantly our GA is able to perform reasonable well on a variety of different types of problems, without code changes and depending only on current the state of the knowledge base.

[3] $iterations = 100$; $populationSize = 10$; $selectionSize = 6$; $tournamentSize = 2$; $tournamentProbability = 0.5$; $growProbability = 0.3$; $shrinkProbability = 0.3$; $mutationProbability = 0.3$

3 Conclusion and Future Work

We have outlined an approach that directly integrates a knowledge representation system and logical reasoning into a mathematical program. This enables us to adhere implicit information during optimization and to model the problem domain more complex and realistic. Our genetic algorithm is able to run on different kinds of problems, without any code changes and by only changing background logic, if the problem structure slightly changes or new domain knowledge becomes present. Furthermore, we are able to define complex knapsack problems directly and ad-hoc on the semantic web, thus avoiding time consuming transformation of information through our query-based approach. In our future work we plan to conduct an in-depth evaluation of performance of this approach in real world problems.

References

1. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2008.
2. P. C. Chu and J. E. Beasley. A Genetic Algorithm for the Multidimensional Knapsack Problem. *J. Heuristics*, 4(1): 63–86, 1998.
3. José E. Gallardo, Carlos Cotta and Antonio J. Fernández. Solving the multidimensional knapsack problem using an evolutionary algorithm hybridized with branch and bound. In José Mira and José R. Álvarez, editors, *IWINAC (2)*, volume 3562 of *Lecture Notes in Computer Science*, pages 21–30. Springer, 2005.
4. Raymond R. Hill and Chaitr Hiremath. Improving genetic algorithm convergence using problem structure and domain knowledge in multidimensional knapsack problems. *International Journal of Operational Research*, 1(1–2): 145–159, 2005.
5. Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, May 2004.
6. Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Springer Berlin Heidelberg, 2 2004.
7. Brian McBride. Jena: A Semantic Web Toolkit. *IEEE Internet Computing*, 6(6): 55–59, 2002.
8. Boris Motik, Ulrike Sattler, and Rudi Studer. Query Answering for OWL-DL with rules. *J. Web Sem.*, 3(1): 41–60, 2005.
9. David Pisinger. The quadratic knapsack problem – a survey. Discrete Applied Mathematics, 155(5): 623–648, 2007.
10. Tugba Saraç and Aydin Sipahioglu. *A genetic algorithm for the quadratic multiple knapsack problem*. In Francesco Mele, Giuliana Ramella, Silvia Santillo, and Francesco Ventriglia, editors, *BVAI*, volume 4729 of *Lecture Notes in Computer Science*, pages 490–498. Springer, 2007.
11. Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3): 96–101, 2006.
12. Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical OWL-DL reasoner. *J. Web Sem.*, 5(2): 51–53, 2007.
13. W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax, Februrary 2004.
14. W3C. OWL 2 Web Ontology Language, October 2009.
15. Du Wei and Li Shuzhuo. An Artificial Intelligence Algorithm for Multi-dimensional Knapsack Problem Based on Small World Phenomenon. *2009 WRI World Congress on Computer Science and Information Engineering*, pages 665–669, March 2009.

A New Representation for the Fuzzy Systems

Iuliana Iatan and Stefan Giebel

Abstract A new representation for fuzzy systems in terms of additive and multiplicative subsystem inferences of single variable is proposed. This representation enables an approximate functional characterization of the inferred output. The form of the approximating function is dictated by the choice of polynomial, sinusoidal, or other designs of subsystem inferences.The first section presents the proposed subsystem inference representation. The next two sections discuss the cases of polynomial, sinusoidal and exponential designs of subsystem inferences and their applications. The section 4 presents some conclusions.

1 Subsystem Inference Representation

Consider a fuzzy system with two input variables a and b with rule consequents embedded in the $m_a \times m_b$ matrix

$$U_{m_a,m_b} = \begin{pmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,m_b} \\ u_{2,1} & u_{2,2} & \dots & u_{2,m_b} \\ \dots & \dots & \dots & \dots \\ u_{m_a,1} & u_{m_a,2} & \dots & u_{m_a,m_b} \end{pmatrix}. \tag{1}$$

The inferred output is [9]

$$\mathcal{I}_{a,b}(U_{m_a,m_b}) = \frac{\mathcal{M}_A^T \cdot U_{m_a,m_b} \cdot \mathcal{M}_B}{\left(\sum_{i=1}^{m_a} \mu_{A_i}(a)\right) \cdot \left(\sum_{j=1}^{m_b} \mu_{B_j}(b)\right)}, \tag{2}$$

Iuliana Iatan
Department of Mathematics and Computer Science, Technical University of Civil Engineering, Bd. Lacul Tei 124, sector 2, 38RO-020396 Bucharest, Romania, e-mail: iuliafi@yahoo.com

Stefan Giebel
Technical University of Braunschweig, Institut für Psychologie Spielmannstr. 12a / 2. OG, D-38106 Braunschweig e-mail: Stefan.Giebel@gmx.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_33, © Springer-Verlag Berlin Heidelberg, 2011

where the column vector $\mathcal{M}_A$ and respectively $\mathcal{M}_B$ contain the membership degrees of a and b:

$$\mathcal{M}_A = \begin{pmatrix} \mu_{A_1}(a) \\ \mu_{A_2}(a) \\ \vdots \\ \mu_{A_{m_a}}(a) \end{pmatrix}, \quad \mathcal{M}_B = \begin{pmatrix} \mu_{B_1}(b) \\ \mu_{B_2}(b) \\ \vdots \\ \mu_{B_{m_b}}(b) \end{pmatrix}. \tag{3}$$

Let a set of linear independent m_a by one column vectors $f_A^{(q)}$, $(\forall)\ q = \overline{1,m_a}$, be selected for variable a and, similarly, a linear independent set of m_b by one column vectors $f_B^{(l)}$, $(\forall)\ l = \overline{1,m_b}$, be selected for variable b. One can associate with each vector $f_A^{(q)} = [f_A^{(q)}(1)\ f_A^{(q)}(2)\dots f_A^{(q)}(m_a)]^T$ a subsystem $A^{(q)}$ with the following fuzzy rules: *Subsystem* $A^{(q)}$: *Rule* i : $A_i \to \delta(f_A^{(q)}(i))$, $(\forall)\ i = \overline{1,m_a}$, where δ is an output membership function. The inferred output of subsystem $A^{(q)}$ becomes [9]:

$$\mathcal{I}_a(f_A^{(q)}) = \frac{\left(\sum_{i=1}^{m_a} f_A^{(q)}(i) \cdot \mu_{A_i}(a)\right)}{\sum_{i=1}^{m_a} \mu_{A_i}(a)} = \frac{\mathcal{M}_A^T \cdot f_A^{(q)}}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}. \tag{4}$$

Similarly, for each column vector $f_B^{(l)} = [f_B^{(l)}(1)\ f_B^{(l)}(2)\dots f_B^{(l)}(m_b)]^T$, one constructs a subsystem $B^{(l)}$ with the fuzzy rules: *Subsystem* $B^{(l)}$: *Rule* j : $B_j \to \delta(f_B^{(l)}(j))$, $(\forall)\ j = \overline{1,m_b}$ and inferred output

$$\mathcal{I}_b(f_B^{(l)}) = \frac{\left(\sum_{j=1}^{m_b} f_B^{(l)}(j) \cdot \mu_{B_j}(b)\right)}{\sum_{j=1}^{m_b} \mu_{B_j}(b)} = \frac{(f_B^{(l)})^T \cdot \mathcal{M}_B}{\sum_{j=1}^{m_b} \mu_{B_j}(b)}. \tag{5}$$

The selection of vectors $f_A^{(q)}$ and $f_B^{(l)}$ should depend on the kind of approximation function one desires for the problem at hand, be it polynomial, sinusoidal, or other designs.

The inferred output $\mathcal{I}_{a,b}(U_{m_a,m_b})$ is now expressible in terms of the inferred outputs $\mathcal{I}_a(f_A^{(q)})$ and $\mathcal{I}_b(f_B^{(l)})$. Using the fact that the $m_a \times m_a$ and $m_b \times m_b$ matrices $F_A = (f_A^{(1)},\ f_A^{(2)}, \dots, f_A^{(m_a)})$ and $F_B = (f_B^{(1)},\ f_B^{(2)}, \dots, f_B^{(m_b)})$ are invertible, (3) becomes:

$$\mathcal{I}_{a,b}(U_{m_a,m_b}) = \left(\frac{\mathcal{M}_A^T \cdot F_A}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}\right) \cdot \nu_{m_a,m_b} \cdot \left(\frac{(F_B)^T \cdot \mathcal{M}_B}{\sum_{j=1}^{m_b} \mu_{B_j}(b)}\right), \tag{6}$$

where the $m_a \times m_b$ matrix ν_{m_a,m_b} is defined by $\nu_{m_a,m_b} \equiv F_A^{-1} \cdot U_{m_a,m_b} \cdot (F_B^T)^{-1}$.

With

$$\left(\frac{\mathcal{M}_A^T \cdot F_A}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}\right) = \left[\frac{\mathcal{M}_A^T \cdot f_A^{(1)}}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}, \frac{\mathcal{M}_A^T \cdot f_A^{(2)}}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}, \dots, \frac{\mathcal{M}_A^T \cdot f_A^{(m_a)}}{\sum_{i=1}^{m_a} \mu_{A_i}(a)}\right] =$$

$$= \left(\mathcal{I}_a(f_A^{(1)}),\, \mathcal{I}_a(f_A^{(2)}),\ldots,\mathcal{I}_a(f_A^{(m_a)})\right)$$

and similarly,

$$\left(\frac{(F_B)^T \cdot \mathcal{M}_B}{\sum_{j=1}^{m_b} \mu_{B_j}(b)}\right) = \begin{pmatrix} \mathcal{I}_b(f_B^{(1)}) \\ \mathcal{I}_b(f_B^{(2)}) \\ \vdots \\ \mathcal{I}_b(f_B^{(m_b)}) \end{pmatrix}$$

(2) becomes

$$\mathcal{I}_{a,b}(U_{m_a,m_b}) = \left(\mathcal{I}_a(f_A^{(1)}),\, \mathcal{I}_a(f_A^{(2)}),\ldots,\mathcal{I}_a(f_A^{(m_a)})\right) \cdot \nu_{m_a,m_b} \cdot \begin{pmatrix} \mathcal{I}_b(f_B^{(1)}) \\ \mathcal{I}_b(f_B^{(2)}) \\ \vdots \\ \mathcal{I}_b(f_B^{(m_b)}) \end{pmatrix} =$$

$$= \sum_{q=1}^{m_a} \sum_{l=1}^{m_b} \nu_{q,l} \cdot \mathcal{I}_a(f_A^{(q)}) \cdot \mathcal{I}_b(f_B^{(l)}), \tag{7}$$

where $\nu_{q,l}$ are elements of the matrix ν_{m_a,m_b}.

To recover U_{m_a,m_b} from ν_{m_a,m_b}, one has $U_{m_a,m_b} = F_A \cdot \nu_{m_a,m_b} \cdot (F_B^T)$.

Equation (5) expresses the original inference in terms of multiplicative subsystem inferences of single variable a and b. A fuzzy system of two input variables is hence generally expressible as additive sum of $m_a \times m_b$ systems each of which is multiplicative decomposable into two subsystems of single variable. The previous result can be extended to a fuzzy system with a general number of n input variables. The inferred output is expressible in the following form:

$$\mathcal{I}_{a,b,c,d,\ldots,x,y,z}(U_{m_a,m_b,m_c,m_d,\ldots,m_x,m_y,m_z}) = \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} \sum_{k=1}^{m_c} \sum_{s=1}^{m_d} \cdots \sum_{t=1}^{m_x} \sum_{h=1}^{m_y} \sum_{r=1}^{m_z} \nu_{i,j,k,s,\ldots,t,h,r} \cdot$$

$$\cdot \mathcal{I}_a(f_A^{(i)}) \cdot \mathcal{I}_b(f_B^{(j)}) \cdot \mathcal{I}_c(f_C^{(k)}) \cdot \mathcal{I}_d(f_D^{(s)}) \cdot \ldots \cdot \mathcal{I}_x(f_X^{(t)}) \cdot \mathcal{I}_y(f_Y^{(h)}) \cdot \mathcal{I}_z(f_Z^{(r)}). \tag{8}$$

Equation (8) expresses a general fuzzy system of n variables into $m_a \times m_b \times \ldots \times m_y \times m_z$ systems each of which is multiplicative decomposable into n subsystems of single variable.

2 Some Designs of Subsystem Inferences

Kosko ([6]), Wang ([8]) and later Alci ([1]), Kim ([5]) have shown that fuzzy systems are universal approximators to continuous functions on compact domain. The same conclusion can be reached here with subsystem inference representation.

A fuzzy system is expressible in terms of polynomial inferences, as [9]:

$$\mathcal{I}_{a,b}(U_{m_a,m_b}) = \nu_{1,1} + \sum_{l=2}^{m_b} \nu_{1,l} \cdot \mathcal{I}_b(f_B^{(l)}) + \sum_{q=2}^{m_a} \nu_{q,1} \cdot \mathcal{I}_a(f_A^{(q)}) +$$

$$+ \sum_{q=2}^{m_a} \sum_{l=2}^{m_b} \nu_{q,l} \cdot \mathcal{I}_a(f_A^{(q)}) \cdot \mathcal{I}_b(f_B^{(l)}). \tag{9}$$

As m_a and m_b tend to large values, the polynomial inferences $\mathcal{I}_a(f_A^{(i)})$ and $\mathcal{I}_b(f_B^{(j)})$ converge uniformly to the polynomial terms $a^{(i-1)}$ and $b^{(j-1)}$. This shows that $\mathcal{I}_{a,b}(U_{m_a,m_b})$ is an approximation to the polynomial function

$$p(a,b) = \nu_{1,1} + \sum_{l=2}^{m_b} \nu_{1,l} \cdot (b-b_1)^{l-1} + \sum_{q=2}^{m_a} \nu_{q,1} \cdot (a-a_1)^{q-1} +$$

$$+ \sum_{q=2}^{m_a} \sum_{l=2}^{m_b} \nu_{q,l} \cdot (a-a_1)^{q-1} \cdot (b-b_1)^{l-1}. \tag{10}$$

Since polynomials are universal approximators (see [7]), the same can be concluded regarding fuzzy system.

The vectors $f_A^{(q)}$, $q = \overline{1, m_a}$ can also be selected to emulate other functions. These vectors give rise to subsystem inferences $\mathcal{I}_a(f_A^{(q)})$ which approximate the sinusoidal terms of $\sin(r \cdot (a-\tilde{a}))$ and $\cos(r \cdot (a-\tilde{a}))$. Same as before for polynomial inferences, the fuzzy inferred output constitutes a piecewise linear approximation of the sinusoidal function; $\mathcal{I}_{a,b}(U_{m_a,m_b})$ will be an approximation to:

1) the sinusoidal function (if m_a is odd):

$$p(a,b) = \nu_{1,1} + \sum_{l=2}^{m_b} \nu_{1,l} \sin\left(\frac{l-1}{2}(b-\tilde{b})\right) + \sum_{q=2}^{m_a} \nu_{q,1} \cdot \sin\left(\frac{q-1}{2}(a-\tilde{a})\right) +$$

$$+ \sum_{q=2}^{m_a} \sum_{l=2}^{m_b} \nu_{q,l} \cdot \sin\left(\frac{q-1}{2} \cdot (a-\tilde{a})\right) \cdot \sin\left(\frac{l-1}{2} \cdot (b-\tilde{b})\right) \tag{11}$$

2) the cosinusoidal function (if m_a is even):

$$p(a,b) = \nu_{1,1} + \sum_{l=2}^{m_b} \nu_{1,l} \cdot \cos\left(\frac{l}{2} \cdot (b-\tilde{b})\right) + \sum_{q=2}^{m_a} \nu_{q,1} \cdot \cos\left(\frac{q}{2} \cdot (a-\tilde{a})\right) +$$

$$+ \sum_{q=2}^{m_a} \sum_{l=2}^{m_b} \nu_{q,l} \cdot \cos\left(\frac{q}{2} \cdot (a-\tilde{a})\right) \cdot \cos\left(\frac{l}{2} \cdot (b-\tilde{b})\right). \tag{12}$$

In the example of exponential subsystem inferences,

$$f_A^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}, \quad f_A^{(2)} = \begin{pmatrix} 1 \\ e^{(a_2-a_1)} \\ e^{(a_3-a_1)} \\ \vdots \\ e^{(a_{m_a-1}-a_1)} \\ e^{(a_{m_a}-a_1)} \end{pmatrix}, \ldots, f_A^{(q)} = \begin{pmatrix} 1 \\ e^{(q-1)\cdot(a_2-a_1)} \\ e^{(q-1)\cdot(a_3-a_1)} \\ \vdots \\ e^{(q-1)\cdot(a_{m_a-1}-a_1)} \\ e^{(q-1)\cdot(a_{m_a}-a_1)} \end{pmatrix},$$

$$\ldots, \ f_A^{(m_a)} = \begin{pmatrix} 1 \\ e^{(m_a-1)\cdot(a_2-a_1)} \\ e^{(m_a-1)\cdot(a_3-a_1)} \\ \vdots \\ e^{(m_a-1)\cdot(a_{m_a-1}-a_1)} \\ e^{(m_a-1)\cdot(a_{m_a}-a_1)} \end{pmatrix}$$

the inference $\mathcal{I}_a(f_A^{(q)})$ constitutes a piecewise linear approximation to the exponential term $e^{(q-1)\cdot(a-a_1)}$.

3 Application

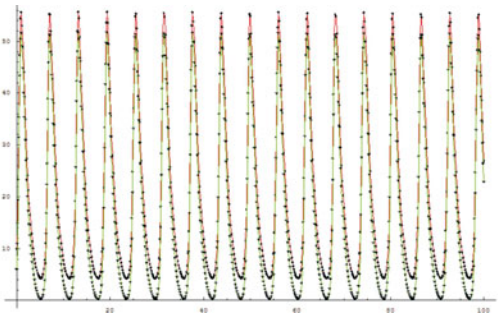
Especially in the case of having no analytical solution the new procedure is interesting. The Lotka-Volterra equations is also called the predator-prey equations. The equations are a pair of first-order, non-linear, differential equations. They are needed to describe the dynamics of biological systems [2] in which two species interact with each other. One is the predator and one its prey. If there are not enough preys, the population of predator will decrease. And if the population of preys is increasing, also the population of predator will increase. It can be also applied in economics. They develop [4] in time according to the pair of equations:

$$\frac{dx}{dt} = x(\alpha - \beta y) \tag{13}$$

$$\frac{dy}{dt} = -y(\gamma - \delta x), \tag{14}$$

where: y (=10) is the number of some predator (for example, lions); x (=800) is the number of its prey (for example, zebras); $\frac{dy}{dt}$ and $\frac{dx}{dt}$ represent the growth of the two populations against time; t represents the time, and $\alpha = 3$, $\beta = 0.1$, $\gamma = 0.8$ and $\delta = 0.002$ are parameters representing the interaction of the two species. The development of each species for a special time interval can be described also by the upper procedure in the form of a polynom. The values for t should be adapted to the procedure. In the next figure it can be seen that the method is usefully to approximate (green) the function of one population(red). The number of prey can be approximated easily by the upper procedure. Furthermore, the Lotka-Volterra equations [3] are used in economy. Similar relations are between different kind of

industries, as an example between engine construction and mining. Furthermore the economic cycle in general can be simulated.



4 Conclusions

A new representation for fuzzy systems in terms of additive and multiplicative subsystem inferences of single variable is proposed. This representation enables an approximate functional characterization of the inferred output. The form of the approximating function is dictated by the choice of polynomial, sinusoidal, or other designs of subsystem inferences. With polynomial subsystem inferences, the approximating function is a sum of polynomial terms of orders depending on the numbers of input membership functions. The constant, linear, and nonlinear parts of the fuzzy inference can hence be identified. The present work also includes an application which shows that the procedure can approximate very well the differential equation. In the case of no analytical solution the procedure is a good alternative.

References

1. M. Alci. Fuzzy rule-base driven orthogonal approximation. *Neural Computing & Applications*, 17(5–6): 501–507, 2008.
2. F. Brauer and C. Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. Springer-Verlag, 2000.
3. G. Gandolfo. Giuseppe Palomba and the Lotka-Volterra equations. *Rendiconti Lincei*, 19(4): 347–357, 2008.
4. R. Haberman. *Mathematical models: mechanical vibrations, population dynamics, and traffic flow*. Verlag SIAM, 1998.
5. J. Kim and J. Lee. Single-input single-output convex fuzzy systems as universal approximators for single-input single-output convex functions. In *Proceedings of the 18th international conference on Fuzzy Systems*, pages 350–355, 2009.
6. B. Kosko. Fuzzy systems as universal approximators. *IEEE Transactions on Computers*, 43(11): 1329–1333, 1994.
7. W. Rudin. *Principles of Mathematical Analysis*. 3rd ed., McGraw-Hill, 2001.
8. L. Wang and J. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least square learning. *IEEE Trans Neural Networks*, 3: 807–814, 1992.
9. Y. Yam. Subsystem Inference Representation for Fuzzy Systems Based Upon Product-Sum-Gravity Rule. *IEEE Transactions on Fuzzy Systems*, 5(1): 90–107, 1997.

Agent-Based Cooperative Optimization of Integrated Production and Distribution Planning

Y. Hu, O. Wendt, and A. Keßler

1 Introduction

The cooperative management of inter-organizational supply chains has attracted a great deal of interest in recent years. To achieve business success, companies should not only focus on individual internal supply chains, but also on the coordination and collaboration of business processes and network-level decision-making along tiers of their supply chain networks (SCN)[1] [12]. The cost savings realized by cooperation between the decision makers is significant in many cases. Since real-life supply chain management is closely related to problems caused by the divergent interests of the actors (enterprises) and the distributed structure of the underlying optimization (scheduling), one natural way to address this constellation is to employ a holonic multi-agent system (MAS), in which the process structure is mapped onto single agents [3]. In this paper we address the question of how production and logistics costs may be reduced by exploiting the scheduling flexibilities of the supply chain partners and employ an integrated approach to coordinate and optimize business processes by synchronizing the interacting production and logistics scheduling across the multi-tier supply web. For the producer and the distributor represented by single agents, various cost and schedule optimization problems are defined.

Consider the scenario: The preliminary or final items must be picked up at the production facility and delivered to one or more different production facilities on the next tier or to the end customers. Obviously, the due time (DT) for a production job of a producer corresponds to the earliest pickup time for producers on subsequent tiers, and the release time (RT) of a producer or an end customer corresponds to the latest delivery time of the supplier for the required preliminary items (or raw materials) from the predecessing producers. The pickup and delivery time intervals are defined by the contracts between the production and logistics/transport agents

Business Information Systems & Operations Research
University of Technology, Kaiserslautern, Germany, e-mail: {hu, kessler, wendt}@wiwi.uni-kl.de

[1] Due to its structural, we term the multi-tier SCN considered in this paper as *supply web*.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_34,

(PA and TA). Thus, we only consider two different agent types in the supply web, namely PA and TA. In addition to the common objective minimizing the total operating and outsourcing costs of the tasks, agents can try to increase their temporal flexibility – and thus the utility of their individual contract situation – by renegotiating for new contract execution time with their contract partners. However, since this relaxation of the agent's own problem leads to a more constrained problem for the contracting agent, compensation payments will have to be negotiated to find an appropriate trade-off as well. The negotiation of contract execution time has to be regarded as an inter-agent schedule optimization process, leading to the social contract equilibrium of the MAS based on prices calculated according to the production and transportation load.

2 Agents' Internal Optimization

The PA's production planning consists of determining a single-machine schedule which minimizes the internal production and outsourcing cost, satisfies the external demands for items, fulfills the resource capacity constraints and complies to the agreed production due time and delivery time of PA's supplier respectively. A simple version of the PA's internal optimization problem is defined as a Weighted Job Interval Scheduling Problem (WJISP) introduced by [5] as an extension and decomposition of JSSP by defining weights to each of jobs. In contrast to DISPOWEB ([13]) where a constant economic weight is assigned to each of the single tasks in a job, a time dependent price function is introduced for the single tasks in the supply web. The time intervals $[RT, DT]$ of a PA's WJISP here are given by the contracts between agents (cp. section 1). We suppose the production costs caused by processing the tasks on the shop to be constant. Then, only the delivery contracts and their outsourcing costs (as defined by the tasks' weights, if they cannot be scheduled within their interval), as well as the sequence-dependent setup times/costs are relevant to the objective function Z_P which is formulated in eq.(1).

$$\text{Min. } Z_P = \sum_{C,C' \in \mathbf{PC}} k_{CC'} X_{CC'} + \sum_{C \in \mathbf{PC}} wp_C (1 - YP_C) \tag{1}$$

for the production contract set **PC** of PA, the sequence-dependent setup costs $k_{CC'}$ occurred by processing C before C' (i.e. the binary sequence variable $X_{CC'} = 1$). wp_C being the outsourcing cost of $C \in \mathbf{PC}$, if C not excutable, i.e. the binary variable $YP_C = 0$.

Each TA represents a fleet of capacitated vehicles and has to plan its optimal route X and schedule S for executing the paired pickup and delivery tasks $C_p \in \mathbf{TC_i}$ and $C_d \in \mathbf{TC_j}$ from agent $i \in A$ to agent $j \in A$ (i.e. the goods picked up at i must be delivered at specified j by the same vehicle). A denotes the set of production agents representing the production plants, and TC_i the set of transportation contracts of TA with $i \in A$, due to the fact that an agent i in general produces as well as requires heterogeneous items simultaneously. Thus at i will be picked up and delivered more

than once according to the contracts with TA. In analogy to the WJISP we introduce the time-dependent price function for the single pickup and delivery tasks of a vehicle's tour and allow for outsourcing transportation tasks by incurring costs wt_{iC} (presented by the tasks' weight as well) whenever the time interval cannot be met. We label this "economic" extension of the Vehicle Routing and Scheduling Problem (cp. e.g. [2]) where the precedence relations between customers to be served are often pre-determined because of exact service time or time intervals given by the customers, the Weighted Pickup and Delivery Interval Routing and Scheduling Problem (WPDIRSP) with the objective (s. eq.(2)) of minimizing the total travel and outsourcing costs formulated as follows, subject to the further restrictions: the coupling constraints between corresponding pickup and delivery contracts, each of contracts with $i \in A$ will be processed once at most.

$$Z_T = \sum_{\mu \in V} \sum_{i,j \in \hat{A}} \sum_{C' \in \mathbf{TC_i}} \sum_{C \in \mathbf{TC_j}} X_{ij\mu}^{C'C} \cdot c_{ij} + \sum_{i \in A} \sum_{C \in \mathbf{TC_i}} wt_{iC}(1 - YT_{iC}) \tag{2}$$

where $\hat{A} := A \cup \{0, n+1\}$ defines the set of the stations to be visited by vehicles $\mu \in V$ (set of vehicles). The start and end depot are represented by the fictive plants 0 and $n+1$ respectively. c_{ij} denotes the travel cost from i to j. The variables $X_{ij\mu}^{C'C}$ and $YT_{iC} \in \{0,1\}$ indicate the executing sequence of contracts and the outsourcing option of vehicle μ.

To facilitate the agents' collaborative optimization as well as to provide the maximum adaptability of our model to real world scenarios, the PAs' planning model is adapted to a dynamic one with a weak degree of dynamism (DoD)[2]. PAs are allowed to negotiate and contract new services or products within the actual planning horizon, if free capacity is available. The TAs operate, however, a dynamic planning with a moderate (or high) degree. The contract allocation between agents occurs in a sealed bid auction (cp. section 3). So the WPDIRSP in supply web can be classified as a partial dynamic VRP (cp. e.g. [6]).

3 Agents' Cooperative Contract Optimization

Cooperative contract optimization can be achieved either by out-/insourcing the task or by (re-)contracting the release time (pickup time) and due time (delivery time) with the contract partners, aiming to maximize their total profits. Due to the distributed structure of the network, we primarily require a network protocol that enables the agents to be aware of the active agents in the network and their scope of services (s. below, directory service). Next, a failsafe, reliable and efficient communication between the agents has to be provided. For this purpose, we designed a slim Distributed Recursive Converging 2-Phase Network Protocol (DRC_2PNP) extending to the Contract Net Protocol (CNP) defined by [1].

[2] That means, more than 80% of customer requests are known at the beginning of planning. The DoD is measured by the number of dynamic requests relative to the total [8]

A request for quotation (RFQ) from contracting initiator defines a service request with desired service execution time and a this time dependent contract price function (optinal), whereas a bid can be regarded as a temporary contract giving the contract execution time (which could deviate from the desired one) determined by internal optimization mechanism and the corresponding price. Its second phase is similar to the third one of the CNP. The communication procedures of the DRC_2PNP differ in two manners. A directory service (DS) is introduced that helps the agents to find the potential business partners, i.e. the DS manages a database with the contact information and the scope of services of all active agents. Furthermore, the agents' online status is tracked by the DS. By means of the agents' contact information gained by the DS, each network member is able to conduct a direct peer-to-peer communication with the other agents. To avoid a pause for bids on a RFQ and the acknowledgment (ACK) of the bid, respectively, two deadlines are introduced for each of the phases and involved in the RFQ sent by the demanding agent. The first deadline D_1 serves as a recursion depth limiter preventing recursive outsourcing cascades. Furthermore, agents do not need to reject uninteresting RFQs explicitly after the expiration of D_1 and the requesting agent will not accept any further bids. Similar to D_1, the use of D_2 avoids an explicit refusal of the received offers. The bidding agents discard the received RFQ and their corresponding bid, if D_2 has expired and no ACK is received. The demanding agent selects the most advantageous one among the received bids by means of the cooperative acceptance criteria based on that of simulated annealing (cp. [4, 13]. Like the trust accounting mechanism, the cooperative acceptance criteria are integrated into the schedule optimization procedure.

4 Preliminary Results

In order to benchmark the agent's distributed planning mechanism and validate the agents' internal cost and schedule optimization models, computational simulation using CPLEX 11.0.0. Due to the lack of benchmark problems for integrated planning and scheduling of SCN, the problem instance sets are generated based on the Solomon's instances R1 and R2 of VRPTW. Unlike [7] who made a random pairing of the customer nodes within routes in solutions obtained by heuristic for VRPTW, and [9] where the customers on the routes in optimal solutions are paired, we generate the contract pairs between customer nodes randomly, so that a node representing the PA in our supply web will be served as a pick up as well as a delivery station and visited repeatedly due to the multiple contracts. Furthermore, we introduce the outsourcing option and outsourcing costs for individual contracts. So our supply web instances represent a more realistic scenario and can be regarded as an extension of dynamic Pickup and Delivery Problem (PDP)[3] with multiple vehicles and heterogeneous goods.

[3] A comprehensive survey of PDP can be found in [11]

For the computational simulation, 80 instances of various network sizes are generated and tested. In view of the fact that both of the optimization problems are NP-hard, small instances are solved by CPLEX but bounds on the globally optimal solution quality are obtained, as shown in the left table in Fig. 1. The computational results of small size supply web instances with one TA operating 4 vehicles capacitated by 250 QU and 20 contracts with 10 PAs are given in the first row of the table. The second row shows the results for instances with 30 contracts and 15 PAs.

Aver. CPU time for 1. feas. sol.	Aver. cost of 1. feas. sol.	Aver. gap to current best bound	Average cost of best sol.	Aver. gap to current best bound
174,28	383,15	61,14%	234,57	23,84%
512,84	491,78	58,37%	357,57	40,56%

Aver. CPU time	Aver. total cost	Aver. outs. cost	# Outs. con. (%)
37,96	5222,74	69,8%	34,5%
73,51	13543,14	94,1%	61,2%

Fig. 1 Simulation Results for Supply Web Instances

Therefore, for both of agents' internal scheduling problems we developed a hybrid meta-heuristic algorithm combining simulated annealing and tabu-search (SATS). We introduce a fictive vehicle v_0 with unlimited capacity for tasks to be outsourced. SATS uses the modified λ-interchange machansim [10] for $\lambda \in \{1,2\}$ to generate neighboring solutions. For a given feasible travel plan X describing the m vehicle's routes X_v ($v = 1,\ldots,m$), the 1-interchange between a pair of routes (X_u , X_v) replaces a subset of contract pairs $X_\mu \subseteq X_u$ ($|X_\mu| \leq 1$) (in case of WJISP a single production contract) by a contract pair $X_v \subseteq X_v$ ($|X_v| \leq 1$) to obtain two new route sets and a new neighboring solution. $(0,1)$ and $(1,0)$ operators shift one pickup and delivery contract pair from one route to another, whereas $(1,1)$ interchange process exchanges or relocates each pickup and delivery contract pair from one route with every contract pair to a different route. Each feasible exchange or relocation of contract pairs between two routes is evaluated by the cost function consisting $f(X,S)$ of the change of travel costs, outsourcing costs and tour duration. In order to avoid moves that result in cycles and also force the exploration of neighborhood, a Tabu-List is introduced to save the information about the move performed in the currently accepted solution: the pair X_μ of pickup and delivery contract removed from X_u, as well as the iteration number, at which the move occurred. So that a reinsertion move of this contract pair into the route X_u by the local search procedure is tabu for a predefined duration. The length of $TL = \{(u,\mu,v,\nu,k)\}$ varies randomly between $(\theta_{\min},\theta_{\max})$ each time the temperature is reset. We employ the first-improvement strategy to select the first feasible in the neighbourhood of initial solution generated by 1-interchanging between a pair of routes and accept this move according to a metropolis probability. The temperature will be updated according to the non-monotonic cooling schedule (cp.[10]).

In the case of dynamic planning, we employed the nearest neighbour policy to schedule the $10\% - 20\%$ contracts which occurred during the planning. The computational results for instances with one TA operating 25 vehicles capacitated by 250 QU and 190 contracts with 100 PAs, are represented in first row of the right table in 1. The second row shows the results for instances with one TA planning for

25 vehicles apacitated by 500 QU and 390 contracts with 100 PAs. Due to the lack of vehicle capacities and the narrow time intervals, lots of contracts are outsourced causing the high costs.

5 Further Research

In further research, we will further investigate more suitable and efficient (meta-) heuristics. Further computational simulations in various supply web scenarios will be performed in order to compare the negotiation mechanism with the one based on trust accounts introduced by [13], as well as explore the agents' behaviour using game theory.

References

1. Randall Davis and Reid G. Smith. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence*, 20: 63–109, 1983.
2. Martin Desrochers, Jan Karel Lenstra, and Martin W. P. Savelsbergh. A classification scheme for vehicle routing and scheduling problems. *European Journal of Operational Research*, 46(3): 322–332, 1990.
3. Torsten Eymann. Co-evolution of bargaining strategies in a decentralized multi-agent system. In *Proceedings of the AAAI Fall 2001 Symposium on Negotiation Methods for Autonomous Cooperative Systems, North Falmouth, MA, November 03–04*, 2001.
4. Andreas Fink. Supply chain coordination by means of automated negotiations. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS '04) – Track 3*, page 30070, Washington, DC, USA, 2004. IEEE Computer Society.
5. M. Garey and D. Johnson. Two-processor scheduling with start-time and deadlines. *SIAM Journal of Computing*, 6: 416–426, 1977.
6. A. Larsen, O. B. G. D. Madsen, and M. Solomon. Partially dynamic vehicle routing – models and algorithms. *Journal of the Operational Research Society*, 38: 637–646, 2002.
7. Haibing Li and Andrew Lim. A metaheuristic for the pickup and delivery problem with time windows. *International Journal on Artificial Intelligent Tools*, 12(2): 173–186, 2003.
8. Karsten Lund, Oli B. G. Madsen, and Jens M. Rygaard. Vehicle routing problems with varying degrees of dynamism. Technical report, Department of Mathematical Modelling, Technical University of Denmark, 1996.
9. William P. Nanry and J. Wesley Barnes. Solving the pickup and delivery problem with time windows using reactive tabu search. *Transportation Research Part B: Methodological*, 34(2): 107–121, 2000.
10. Ibrahim Hassan Osman. Metastrategy simulated annealing and tabu search algorithms for the vehicle routing problem. *Ann. Oper. Res.*, 41: 421–451, 1993.
11. Sophie Parragh, Karl Doerner, and Richard Hartl. A survey on pickup and delivery problems. *Journal für Betriebswirtschaft*, 58(2): 81–117, 2008.
12. James B. Rice and Richard M. Hoppe. Network master & three dimensions of supply network coordination: An introductory essay. Technical report, Cambridge, 2002.
13. Tim Stockheim, Oliver Wendt, and Michael Schwind. A trust-based negotiation mechanism for decentralized economic scheduling. In *Proceedings of the 38th Hawaiian International Conference on System Sciences(HICSS-38), Hilton Waikoloa Village, Big Island, Hawaii*, 2005.

Personnel Rostering by Means of Variable Neighborhood Search

Martin Josef Geiger

Abstract The article presents a Variable Neighborhood Search approach for personnel rostering problems, with a particular computer implementation (and results) for the data sets of the very recent First International Nurse Rostering Competition 2010. In this context, our work is motivated by two objectives: (i) The very fast computation of qualitatively good solutions, and (ii) the subsequent provision of decision aid by means of computational intelligence techniques. Starting from initially constructed solutions, we are able to demonstrate a considerable improvement of the quality of the solutions by means of the proposed metaheuristic. Moreover, and due to the rather problem-independent character of the approach, problem formulations with varying characteristics are equally supported by our solution approach.

1 Introduction and Problem Description

Personnel rostering problems [4] describe a class of important planning problems with applications in many business areas. They are basically characterized by situations in which a set of employees must be assigned to shifts such that some desirable properties are present in the constructed plan. Such properties are commonly categorized into the following two classes.

1. On the one hand, solutions must be *feasible*. Important examples include the minimum coverage of shift during certain times, the consideration of legal requirements, etc.
2. On the other hand, desirably properties reflecting the *quality* of the plan should be present. Typical examples are the consideration of day-off-requests by the staff, particular working patterns, and similar.

Martin Josef Geiger
Helmut-Schmidt-University, University of the Federal Armed Forces Hamburg, Logistics Management Department, Holstenhofweg 85, 22043 Hamburg, Germany, e-mail: m.j.geiger@hsu-hh.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_35,

While it has been pointed out before, that rostering problems are of multi-objective nature [6], many solution approaches combine the desirable properties into a single overall penalty function, which is subsequently minimized. Due to the complexity of most problems from this area, *NP*, (meta-)heuristics play an increasingly important role in this context.

Research into rostering problems has very recently been stipulated by the First International Nurse Rostering Competition (NRC) [3]. The objective of the competition consisted in the proposition and implementation of solution approaches for newly released nurse rostering data sets. Benchmark instances are split into three categories: 'sprint' data sets, which must be solved within little time (around ten seconds), 'medium' instances with around 10 minutes available, and 'long' data sets for which the computation time was around 10 hours per instance. From our perspective, the 'sprint' instances are of particular interest, as they call for solution approaches that allow a considerable fast problem resolution, thus leading to (almost) real-time computational intelligence strategies.

The common problem description of the considered nurse rostering problem consists in the assignment of given amount of nurses to each shift within a given planning horizon. Feasible solutions require, that each nurse is assigned to at most one shift per day. Other aspects are integrated by penalty functions, e. g. the consideration of a minimum or maximum number of assignments of each nurse, the integration of day-off- or shift-off-requests, or the avoidance of certain working patterns. Moreover, nurses are, depending on their qualifications, categorized into (regular) 'nurses' and 'head nurses', and shift-dependent qualifications have to be respected. Altogether, these aspects define the quality of each feasible solution, thus leading to a combinatorial optimization (minimization) problem.

2 Solution Approach

The algorithm used for solving the nurse rostering problem consists of two phases: (i) An initial constructive approach, meeting all shift requirements while respecting the hard constraints, thus leading to a first feasible solution, and (ii) an iterative improvement phase based on local search.

2.1 Constructive Approach

We first assign the number of required nurses to all shifts. In this procedure, shifts are selected such that the ones with the least number of assignable nurses are selected first, thus following the most-constrained-first principle. Any difference between the shifts is due to the given qualification constraints, which in some cases call for 'head nurses' instead of regular nurses only. In case of the availability of a sufficient number of head nurses, the 'alternative skill'-soft constraint is met au-

tomatically in this procedure. Besides, we respect the 'single assignment per day'-constraint, and thus obtain a feasible first assignment.

The constructive procedure leads to a first feasible solution within a few milliseconds. On the other hand, the quality of the first assignment with respect to the soft constraints (penalties) is commonly not very high, which is why an iterative improvement phase follows.

2.2 Iterative Improvement

Motivated by the results presented in [2], an algorithm based on the principles of Variable Neighborhood Search (VNS) [5] has been implemented for the improvement of the initially generated alternative. Our VNS-approach makes use of four neighborhoods, one of which implements a perturbation (shaking) of alternatives once a local optimum with respect to the three other neighborhood operators is reached:

- $\mathbf{N}_1$: For any nurse-shift-assignments: Replace the nurse with another free nurse on that day.
- $\mathbf{N}_2$: For any day: Swop the shift assignments of any pair of the nurses assigned to shifts on that day, except if they are assigned to the same shift.
- $\mathbf{N}_3$: For any pair of days: Swop the shift assignments of any pair of nurses while maintaining feasibility with respect to the 'single assignment per day'-constraint.
- $\mathbf{N}_{perturb}$: For a set of s randomly chosen shifts: Remove all nurses assigned to these s shifts and randomly reassign nurses such that the number of required nurses per shift is met again.

Algorithm 1 states the pseudo-code of our approach. The neighborhoods are searched in order of $\mathbf{N}_1 \rightarrow \mathbf{N}_2 \rightarrow \mathbf{N}_3$, returning to $\mathbf{N}_1$ once $\mathbf{N}_2$ or $\mathbf{N}_3$ manage to improve the current solution. Besides, we search the neighborhoods in a first-best-move-fashion, immediately applying improving moves. In addition to that, the first checked move (position) in the neighborhood is chosen at random. Consequently, the algorithm involves a considerable amount of randomization, and independent runs are expected to lead to different results.

3 Experiments and Results

3.1 Experimental Setting

Respecting the maximum running times as given by the benchmark program of the NRC, we ran our software for 10s on a single core of an Intel Q9650 (3.00 GHz) processor for the 'sprint' instances. Due to the stochastic nature of the solver,

Algorithm 1 Variable Neighborhood Search with Shaking

Require: $\mathcal{N} = \{\mathbf{N}_1, \ldots, \mathbf{N}_K, \mathbf{N}_{perturb}\}$
Compute some first initial solution x; $x_{best} = x$, $k = 1$
repeat
  repeat
    Compute neighborhood $\mathbf{N}_k(x)$
    Select $x' \in \mathbf{N}_k(x)$ such that $\nexists x'' \in \mathbf{N}_k(x) \mid f(x'') < f(x')$
    if $f(x') < f(x)$ **then**
      $x \leftarrow x'$
      $k = 1$
      if $f(x) < f(x_{best})$ **then**
        $x_{best} \leftarrow x$
      end if
    else
      $k \leftarrow k + 1$
    end if
  until $k > K$
  $x \leftarrow \mathbf{N}_{perturb}(x)$
until Termination criterion is met

100 independent test runs have been carried out per instance (keeping the random number seeds for reproduction).

The $\mathbf{N}_{perturb}$-operator makes use of an additional parameter, the number of reassigned shifts s. Values of $s = 2, 4, 6, 8, 10, 15$ have been tested for the 'sprint'-data sets, and, based on the results of these initial test runs, s has been fixed to $s = 6$.

3.2 Results

The following Table 1 reports the results obtained through the constructive and the iterative phase, and compares the values to recently published best known upper and lower bounds [1]. While it is possible to see that feasible solutions are easy to find, the overall penalties after the constructive phase turn out to be rather unsatisfactory. This however must be expected, simply as no particular consideration of the penalties is done in this phase of the solution approach.

On the other hand, the Variable Neighborhood Search approach is able to significantly improve the initially constructed solutions. With respect to the early published 'sprint'-instances, six out of ten instances are solved to optimality. For instance `sprint05`, another best-known solution is found, and the results of the remaining three instances show comparably small deviations to the best known solutions (+1 penalty point). The direct comparison of the improvement phase to the construction phase also proves to give a statistically significant improvement, which is however, due to the nature of the construction algorithm, not surprising.
When analyzing the standard deviations of the results, it becomes clear, that the final results of the early published 'sprint'-instances are rather close together, indicating a certain reliability of the solution approach.

Table 1 Experimental results. Optimal solutions are highlighted with the symbol ***, best known alternatives with *. Other results show their deviation to the best known solution in penalty points.

	Construction		Improvement (VNS)				
Instance	Aver. pen.	Std.dev.	Min. pen.	Aver. pen.	Std.dev.	Best LB	Best UB
1. sprint01	207	9.86	56 (***)	59.23	1.32	56.00	56
2. sprint02	210	9.65	58 (***)	61.61	1.42	58.00	58
3. sprint03	206	9.80	51 (***)	54.91	1.57	51.00	51
4. sprint04	216	10.07	60 (+1)	63.95	1.75	58.50	59
5. sprint05	214	11.37	58 (*)	60.12	1.09	57.00	58
6. sprint06	198	9.21	54 (***)	57.18	1.16	54.00	54
7. sprint07	208	10.48	57 (+1)	60.04	1.51	56.00	56
8. sprint08	202	9.72	56 (***)	58.76	1.33	56.00	56
9. sprint09	213	9.66	55 (***)	58.86	1.55	55.00	55
10. sprint10	209	11.56	53 (+1)	55.91	1.40	52.00	52
11. sprint_late01	262	14.59	41 (+4)	47.32	2.73	37.00	37
12. sprint_late02	244	15.10	44 (+2)	48.69	2.51	41.40	42
13. sprint_late03	268	14.51	51 (+3)	56.97	2.63	47.83	48
14. sprint_late04	1243	68.57	86 (+13)	107.04	8.14	72.50	73
15. sprint_late05	253	13.44	47 (+3)	52.49	2.55	43.67	44
16. sprint_late06	222	13.34	43 (+1)	45.96	1.55	41.50	42
17. sprint_late07	1010	72.46	46 (+4)	57.84	5.11	42.00	42
18. sprint_late08	1038	82.82	17 (***)	21.29	4.08	17.00	17
19. sprint_late09	1179	90.59	17 (***)	20.95	3.20	17.00	17
20. sprint_late10	1008	72.17	47 (+4)	59.26	5.82	42.86	43

The data sets which have been made available at a later time (sprint_late) unfortunately show a different picture. Only two instances are solved to optimality, and the deviations for the other eight data sets are considerably higher. Still, VNS significantly improves upon the results of the constructive phase, but converges to solutions of a qualitatively lower lever. Also, the standard deviation of the results turns out to be much higher as in case of the first ten instances, which indicates that the reliability of the best obtained values decreases. Clearly, our solution approach is less adapted to these data sets.

Interestingly, the properties of the sprint_late-instances differ from the first ten data sets. While all data sets consider the solution of problems with ten nurses over a 28-day period, some of the 'late' instances have more unwanted working patterns which seem to complicate their solution using the here presented approach. This is especially true for sprint_late04. The data sets sprint_late08 and sprint_late09 on the other hand, which are solved to optimality, possess no unwanted working patterns and no shift-off or day-off-request.

As a more technical remark, we may mention that the implementation is able to quickly generate and evaluate solutions, with typical values of around 175 evaluations per millisecond (on the hardware mentioned above). Such numbers may be reached by use of 'delta'-evaluating moves, which only compute the difference induced by a particular neighborhood operator, leaving the rest of the solutions un-

touched. In it's current implementation, which is done in .NET, we measured a speedup of this technique over the complete evaluation of around factor 5.96. For the assignment of employees to shifts, the program makes extensive use of the (.NET-) Dictionary-data type, a data structure in which retrieving values may be done close to $O(1)$.

4 Conclusions

A study of Variable Neighborhood Search for the personnel rostering, with a particular implementation for the First International Nurse Rostering Competition 2010 has been presented. Starting from initially randomly constructed solutions, considerable improvements have been observed by means of this metaheuristic. A fast implementation of such techniques is possible when making use of 'delta-evaluations'. Especially with respect to an interactive setting, where only comparable little time is available for computing alternatives, we may expect this to play an important role.

In comparison to recently reported results [1], the approach yields optimal or at least satisfactory results for large part of the benchmark instances. Several other instances however pose difficulties for the presented concept. In conclusion, we are able to postulate that the results seem to depend on the underlying structures of the data sets. Especially unwanted shift patterns present a complicated structure. Clearly, more fine-tuned and thus superior heuristics/ optimization algorithms can be found for these data sets. On the other hand however, it is worth to mention, that our proposed techniques are fast and mostly problem-independent, and therefore can be expected to present an interesting starting point for future research.

References

1. Edmund K. Burke and Timothy Curtois. An ejection chain method and a branch and price algorithm applied to the instances of the first international nurse rostering competition, 2010. In *Proceedings of the 8th International Conference on the Practice and Theory of Automated Timetabling PATAT 2010*, Belfast, Northern Ireland, August 2010. Queen's University Belfast.
2. Edmund K. Burke, Timothy Curtois, Gerhard Post, Rong Qu, and Bart Veltmann. A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem. *European Journal of Operational Research*, 188: 330–341, 2008.
3. De Causmaecker. Website of the first international nurse rostering competition. http://www.kuleuven-kortrijk.be/nrpcompetition, 2010.
4. A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(1): 3–27, 2004.
5. Pierre Hansen and Nenad Mladenović. Variable neighborhood search: Principles and applications. *European Journal of Operational Research*, 130: 449–467, 2001.
6. Panta Lučic and Dušan Teodorovic. Simulated annealing for the multi-objective aircrew rostering problem. *Transportation Research Part A: Policy and Practice*, 33(1): 19–45, 1999.

II.1 Traffic, Transportation and Logistics

Chair: Prof. Dr. M. Grazia Speranza (Università degli Studi di Brescia, Italy)

We solicit original contributions on phenomena associated with all modes of transportation, present and prospective, including planning, design, economic and operational aspects. We are interested to fundamental theories, coupled with observational and experimental studies of transportation and logistics phenomena and processes, mathematical models, advanced methodologies and novel applications in transportation and logistics systems analysis, planning and design. We intend to cover a broad range of topics that include vehicular and human traffic flow theories, models and their application to traffic operations and management, strategic, tactical and operational planning of transportation and logistics systems; performance analysis methods and system design and optimization; theories and analysis methods for network and spatial activity interaction, economics of transportation system supply and evaluation; methodologies for analysis of transportation user behaviour and the demand for transportation and logistics services. Contributions to new ideas and challenging topics, such as for example optimization-based integrated planning, real time problems, optimization for recovery and disruption handling, are most welcome.

Graph Sparsification for the Vehicle Routing Problem with Time Windows

Christian Doppstadt, Michael Schneider, Andreas Stenger, Bastian Sand, Daniele Vigo, and Michael Schwind

1 Introduction

The Vehicle Routing Problem with Time Windows (VRPTW) is one of the most important and widely studied NP-hard combinatorial optimization problems in the operations research literature. The problem calls for the determination of a minimum-cost set of routes for a fleet of identical vehicles with limited capacities to serve a set of customers that have a given demand and an associated time window in which they can be visited. Due to its computational complexity, VRPTW can only be solved by exact methods for instances of moderate size [2]. However, a large number of successful metaheuristic solution methods have been proposed, which are able to produce high-quality solutions for reasonably-sized instances in limited time. For an extensive review, the reader is referred to [5, 3].

Granular tabu search, introduced by [12] for the classical VRP and the distance constrained VRP, is an extension of the classical tabu search metaheuristic (see [6]). It is based on the use of drastically restricted neighborhoods, called granular neighborhoods, which involve only moves that are likely to belong to good feasible solutions. Due to the significantly lower number of possible moves to evaluate, granularity-based algorithms can preserve the overall solution quality while strongly reducing computation times. Motivated by these results, we investigate the usage of sparse graphs (that are obtained from the original one by removing unpromising arcs) within VRPTW solution methods.

In Section 2, we describe the preprocessing steps used to remove infeasible edges from the original graph. Moreover, we introduce three possibilities for sparsifying

Christian Doppstadt, Andreas Stenger, Michael Schwind
IT-based Logistics, Goethe University Frankfurt e-mail: doppstadt|stenger|schwind@wiwi.uni-frankfurt.de

Michael Schneider, Bastian Sand
BISOR, University of Kaiserslautern e-mail: schneider|sand@bisor.de

Daniele Vigo
DEIS, University of Bologna e-mail: daniele.vigo@unibo.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 227
DOI 10.1007/978-3-642-20009-0_36,

problem graphs based on the study of a series of near-optimal solutions to the classical Solomon VRPTW instances. In Section 3, we investigate the impact of utilizing the resulting sparse graphs within an Ant Colony Optimization (ACO) method for VRPTW on the method's solution behavior.

2 Generation of Sparse Graphs for VRPTW

A sparse VRPTW problem graph is constructed by removing infeasible and unpromising arcs from the original problem graph. The goal is to enable heuristics that operate on the sparse graph to obtain solutions more quickly without strong negative influence on the solution quality. For constructive heuristics, this means a reduction of the number of possible insertion candidates in each steps, for local search methods a reduction of the size of the neighborhood to evaluate. After a short description of some preprocessing steps that eliminate infeasible arcs (Section 2.1), we present three methods for sparsifying the problem graph in Section 2.2.

2.1 Preprocessing

The demand d_i of a customer i has to be delivered within time window $TW = [a_i, b_i]$, where a_i denotes the earliest time of delivery and b_i the latest. If a vehicle arrives at customer i before a_i, it has to wait. The service time for serving customer i is given by s_i. The vehicle capacity is limited to C units and every vehicle has to return to the depot before the depot deadline b_0. The travel time from i to j is c_{ij}. Vehicle capacity, customer time windows and depot deadlines are used as decision criteria to eliminate infeasible arcs, resulting in a sparse c_{ij} matrix. To this end, several well-known rules can be applied (cp. [7, 8]):

$$d_i + d_j > C \qquad \textit{Capacity violation} \tag{1}$$

$$a_i + s_i + c_{ij} > b_j \qquad \textit{Customer time windows} \tag{2}$$

$$(a_i + s_i) + c_{ij} + s_j + c_{j0} > b_0 \qquad \textit{Depot deadlines} \tag{3}$$

Moreover, we narrow time windows of customers by considering travel times to reach and leave them:

$$a_i + s_i + c_{ij} < a_j \quad \forall j \quad \rightarrow \quad a_i' = min[a_j] - c_{ij} - s_i \tag{4}$$

$$b_i + s_i + c_{ij} > b_j \quad \forall j \quad \rightarrow \quad b_i' = max[b_j] - c_{ij} - s_i \tag{5}$$

Time windows narrowing and Rules (2) and (3) are repeatedly applied until no more modifications occur.

2.2 Sparsification Methods

Computation times of many solution methods may be significantly reduced by ignoring unpromising arcs. For CVRP, [12] show that high-quality solutions mainly consist of short arcs. However, this knowledge is not necessarily transferable to VRPTW since tight time window constraints might require the inclusion of long arcs to obtain feasible solutions. In this section, we present three methods for sparsifying a VRPTW problem graph based on different "cost" measures: 1) distance, 2) time window adjusted distance and 3) reduced costs of a corresponding network relaxation. For all three methods, the sparse graph is obtained by keeping only those arcs whose "cost" value is below a certain threshold and all the arcs between customers and the depot (cp. [12]).

2.2.1 Distance

Due to the existence of time windows, one expects a much lower correlation of arc length with good solutions in VRPTW than in CVRP. We analyze several near-optimal solutions of a representative sample of Solomon VRPTW instances, for which not only objective function values but also the routing solutions are available from the literature [1].

Table 1 reports minimum, maximum and average arc length of the entire graph and those of the reported solutions. The letter n denotes the number of customers, K the number of vehicles available. The solution value is given by z^*. The results clearly show that VRPTW solutions also consist mostly of short arcs. Moreover, the average arc length of the solutions are quite similar to those of CVRP instances [12].

Table 1 Comparison of arc length of high-quality solutions to arc length of total graph

				c_{ij} Graph			c_{ij} Solution		
Problem	n	K	z^*	Min	Max	Average	Min	Max	Average
C104	100	10	824.78	1.00	96.18	39.47	1.00	47.43	7.50
C207	100	3	588.29	2.00	96.18	40.37	2.00	20.62	5.71
R107	100	12	1159.84	1.41	91.83	33.95	2.24	34.21	10.55
R108	100	9	980.95	1.41	91.83	33.95	2.24	25.50	9.00
R109	100	11	1235.68	1.41	91.83	33.95	2.00	33.97	11.13
R201	100	4	1281.58	1.41	91.83	33.95	2.24	35.90	12.32
R211	100	2	949.49	1.41	91.83	33.95	2.00	38.18	9.31
RC105	100	13	1733.56	1.00	101.21	44.74	2.00	52.20	15.34
RC106	100	12	1384.93	1.00	101.21	44.74	1.00	50.29	12.37
RC208	100	3	833.97	1.00	101.21	44.74	1.00	30.41	8.10
Average				*1.31*	*95.51*	*38.38*	*1.77*	*36.87*	*10.13*

2.2.2 Time Window Adjusted Distance

The first sparsification approach neglects any information about time windows when determining the set of unpromising arcs. However, the quality of an arc is strongly influenced by time windows of its endpoints. The travel time c_{ij} between two customers i, j is a misleading decision value if the time span between the latest departure time at customer i and the earliest delivery time of customer j is larger than c_{ij}, because then the waiting time w_{ij} is ignored. For this reason, we introduce a new travel time matrix c'_{ij} that implicitly considers time windows. For the case mentioned above, c'_{ij} denotes the difference between $(b_i + s_i)$ and a_j, otherwise c'_{ij} is the travel time c_{ij}, i.e. $c'_{ij} = Max[(a_j - (b_i + s_i)); c_{ij}]$.

Thus, the c'_{ij}-matrix provides the minimum total time between deliveries at customer i and j. Using this approach allows us to implicitly consider time windows in the removal of unprofitable arcs. Similarly to the distance-based approach described above, we analyze the c'_{ij} values of high quality solutions compared to the values of the entire graph (see Tab. 2). The maximal arc values of the graph significantly increase due to the time windows. However, maximal and average values of solutions remain almost on the same level which indicates that customer sequences with long waiting times are rare in high-quality solutions.

2.2.3 Reduced-Cost of Network Relaxation

As proposed by [12], we incorporate reduced cost information to design a rule for sparsifying the problem graph. To this end, we solve the network relaxation of VRPTW as described in [4] by means of an appropriate assignment problem and use it to determine arc reduced costs. The extended cost matrix is very similar to the one described in [11], however, to incorporate the time window aspect, we heuristically use the c'_{ij} defined in the previous section.

Table 2 Comparison of c'_{ij} values of high-quality solutions to those of the total graph

				c'_{ij} Graph			c'_{ij} Solution		
Problem	n	K	z^*	Min	Max	Average	Min	Max	Average
C104	100	10	824.78	1.00	749.00	43.32	1.00	47.43	7.78
C207	100	3	588.29	2.00	2678.00	268.63	2.00	20.62	5.71
R107	100	12	1159.84	1.41	122.00	35.01	2.24	34.21	10.55
R108	100	9	980.95	1.41	122.00	34.23	2.24	25.50	9.00
R109	100	11	1235.68	1.41	91.83	34.33	2.00	33.97	11.13
R201	100	4	1281.58	1.41	667.00	105.27	2.24	94.00	15.85
R211	100	2	949.49	1.41	91.83	33.95	2.00	38.18	9.31
RC105	100	13	1733.56	1.00	165.00	46.33	2.00	52.20	15.56
RC106	100	12	1384.93	1.00	101.21	45.10	1.00	50.29	12.37
RC208	100	3	833.97	1.00	101.21	44.74	1.00	30.41	8.10
Average				*1.31*	*489.01*	*69.10*	*1.77*	*42.68*	*10.54*

3 Numerical Studies

To test the influence of the presented sparsification methods on the solution quality and the runtime of metaheuristics, we use an ACO method for VRPTW that operates on the resulting sparse graph. The ACO algorithm was originally proposed to solve a VRPTW with driver-specific travel and service times. The ACO procedure, the complementary local search method and all parameter settings are described in [9, 10]. In the ACO method, a sparsified graph reduces the number of edges that an ant has to consider as possible next step. For the complementary local search method, we use a straightfoward implementation. We check whether a relocate operator involves an "infeasible" edge. If it does, the move is not carried out. We are aware that further significant speedups can be obtained by explicitly taking into account the graph sparsification as in [12]. For all sparsification methods, we use a granularity threshold that reduces the number of remaining arcs to a given value. This parameter is set to 25% for the conducted tests. All experiments are run on a desktop computer at 2.66 GHz and 4 GB of RAM. For our test, we use the representative sample of Solomon VRPTW instances described above.

The preprocessing steps yielded some quite surprising results: For the tested instances, between 0.09% and 48.15% of edges are infeasible, averaging 24.34%. Subsequently, we conducted two ACO runs for each sparsification method and instance. Table 1 reports the gap (in %) between the distance and the runtime of the best run with the respective sparse graph and the best solution of two runs using the standard graph. All three sparsification methods achieve significant speedup with only very minor losses in solution quality. Concerning the solution quality, the reduced cost-based sparsification method gets the best results, concerning the speedup, the time window adjusted distance-based method shows the best performance. For one instance, our ACO method was not able to find a valid solution with the given number of vehicles, which is indicated by dashes in the table.

4 Conclusion

We presented three methods for sparsifying a VRPTW problem graph. All methods were clearly able to speedup the computation while the solution quality only decreased very slightly. In the future, we plan to evaluate the methods on a larger benchmark set and study the influence of the sparsification threshold on solution quality and runtime. Moreover, we intend to integrate the intensification/diversification methods proposed in [12].

Table 3 Influence of different sparsification methods on ACO performance

	Distance		Distance/TW		Reduced Cost	
Problem	ΔDist	ΔRT	ΔDist	ΔRT	ΔDist	ΔRT
C104	4.22	-30.24	0.11	-32.59	0.25	36.51
C207	0.00	-74.56	0.00	-81.14	0.00	-75.21
R107	-0.07	-21.55	-0.89	-28.63	-0.73	-6.93
R108	1.90	-27.56	1.72	-23.19	0.10	-17.09
R109	-3.41	-11.26	-1.98	-9.14	1.28	-20.14
R201	8.46	-76.24	7.59	-81.53	1.88	-77.06
R211	-0.84	-75.15	1.46	-79.07	1.73	-76.17
RC105	–	–	–	–	–	–
RC106	0.37	-23.94	-1.44	-24.28	-2.29	-9.86
RC208	1.56	-75.19	3.28	-74.45	-1.30	-74.11
Average	1.35	-46.19	1.09	-48.21	0.10	- 43.68

References

1. R. Baldacci. Personal Communication, 2010.
2. R. Baldacci, E. Bartolini, A. Mingozzi, and R. Roberti. An exact solution framework for a broad class of vehicle routing problems. *Computational Management Science*, 7(3): 229–268, 2010.
3. Olli Bräysy and Michel Gendreau. Vehicle routing problem with time windows, Part II: Metaheuristics. *Transportation Science*, 39(1): 119–139, 2005.
4. J. F. Cordeau, G. Desaulnier, J. Desrosiers, M. Solomon, and F. Soumis. The VRP with time windows. In Paolo Toth and Daniele Vigo, editors, *The vehicle routing problem*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.
5. Michel Gendreau, Jean-Yves Potvin, Olli Bräysy, Geir Hasle, and Arne Løkketangen. Metaheuristics for the vehicle routing problem and its extensions: A categorized bibliography. In Bruce L. Golden, editor, *The Vehicle Routing Problem: Latest Advances and New Challenges*, chapter 1, pages 143–169. Springer, 2008.
6. Fred Glover and Fred Laguna. *Tabu Search*. Kluwer, 1997.
7. M. W. P. Savelsbergh. Local search in routing problems with time windows. *Annals of Operations Research*, 4(1): 285–305, 1985.
8. M.W.P. Savelsbergh. An efficient implementation of local search algorithms for constrained routing problems. *European Journal of Operational Research*, 47(1): 75–85, 1990.
9. Michael Schneider, Christian Doppstadt, Bastian Sand, Andreas Stenger, and Michael Schwind. A vehicle routing problem with time windows and driver familiarity. In *Seventh Triennial Symposium on Transportation Analysis*, Tromso, Norway, 2010.
10. Michael Schneider, Christian Doppstadt, Andreas Stenger, and Michael Schwind. Ant colony optimization for a stochastic vehicle routing problem with driver learning. In *IEEE Congress on Evolutionary Computation*, Barcelona, Spain, 2010.
11. Paolo Toth and Daniele Vigo, editors. *The vehicle routing problem*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.
12. Toth, Paolo and Vigo, Daniele. The granular tabu search and its application to the vehicle-routing problem. *INFORMS Journal on Computing*, 15(4): 333–346, 2003.

A New Model Approach on Cost-Optimal Charging Infrastructure for Electric-Drive Vehicle Fleets

Kostja Siefen, Leena Suhl, and Achim Koberstein

Abstract Electric-Drive Vehicle Fleets need proper charging infrastructure available in the area of operation. The number of potential locations can be quite large with differing costs for preparation, installations, operation and maintenance. We consider a planning approach based on vehicle movement scenarios. Given a large set of vehicle movements and parking activities, we search for a cost minimal subset of all charging locations subject to proper supply of all vehicles.

1 Introduction

We consider the perspective of a future operator of an electric drive vehicle fleet. Vehicle engines in our case are solely powered by a permanently installed battery (see [2] for a thorough analysis of recent technologies). The vehicles will be used to serve different mobility demand by the vehicle users, e.g. transport of persons or goods. Movements happen one-way, vehicle drivers may change every time and the set of allowed parking positions can be quite large in urban areas. Fleet characteristics such as the distribution of driving distances, parking positions (and durations) can be completely different between problem instances (see [7]).

To supply energy to all fleet vehicles, a proper charging infrastructure has to be available (see [5, 6]). This is a set of locations with parking space and a technical installation suitable to transfer into the energy carrier of the vehicle. Charging is done with a certain electric power - agreed upon between vehicle and charging spot. The total available power can be different among locations. We do not distinguish between conductive (plug) and inductive (contactless) charging, vehicles just need to stay parked. Nevertheless charging can be started at any given battery level and

Decision Support & Operations Research Lab,
University of Paderborn, Warburger Str. 100, 33098 Paderborn, Germany
e-mail: siefen,suhl,koberstein@dsor.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_37,

may as well be interrupted at any time. In our case, we do not consider the flow of electric energy back into the power grid (see [10] for this concept).

A set of potential locations is prepared in advance, and its size is usually quite large, while only a small subset may be needed. Some locations may already be available while others are mandatory to be built. The planner chooses a subset of all available locations, that means a simultaneous decision how many locations are to be opened, where the stations will be located and which capacity is installed.

2 Problem Description

The choice of charging locations needs to provide enough capacity for all vehicle demands while total costs for installation and operation are to be minimized. The maximum accepted distance between location and vehicle parking position is a planning parameter. If locations are close enough to the vehicle, it can charge its battery with a location-specific power depending on the available parking time. The needed on-site capacity (i.e. total number of available chargers) for each location depends on the amount of simultaneous parking vehicles. Vehicle parking durations may in general be shorter or longer than the time needed to fully charge the battery. Vehicles may have different movement and parking characteristics (i.e. trip distances and parking durations).

Location planning problems in literature (see [3] for a recent survey) usually consider spatially fixed demand points (customers) with a certain need of goods over time. The demand is fulfilled by assigning sufficient quantities to connections between supply points (production sites, warehouses) and consumers. The total costs are derived from the locations investments, the building of capacity and the prices for transportation. These models do not consider the movement of demand points which consume energy.

Coverage problems like ambulance location models (see [1]) choose a cost-minimal set of locations which assure proper response times for the population. Problems like p-center, p-median and MCLP maximize the total coverage for a fixed amount of p locations (see [9]). In our problem domain, we do not have to cover all demand points (vehicle positions). In fact, with small distances and sufficient energy, vehicles may park at many places without charging. Furthermore, the location decision has an impact on the demand. The availability of charging locations in one area can avoid their necessity in other places.

To account for the given aspects, we developed a new model. As planning base we consider a movement scenario of all electric vehicles. Starting from a reference point zero in time, we have the movements of all vehicles with parking activities between. We call this alternate chain of driving and parking of one vehicle a *movement sequence* which consists of many consecutive *sequence elements*. Each sequence element contains the information of parking time and potential supply locations nearby. The possible amount of energy consumption is derived from the parking

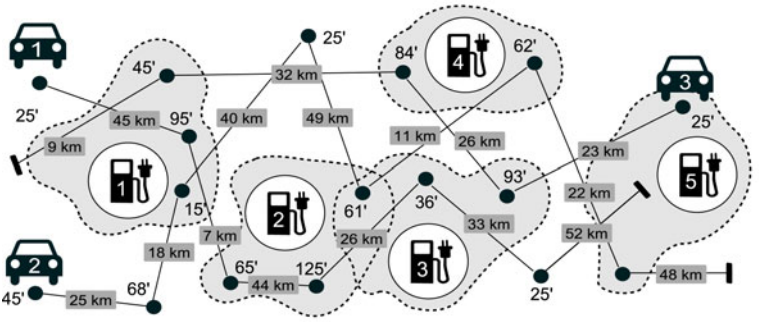


Fig. 1 Illustration of a simple problem instance with 3 vehicles and 5 potential locations. The black points are parking positions with a given parking time. Parking positions can be in range of one or more supply locations (gray areas). Between parking positions, vehicles move a given distance.

time and technological parameters of location and vehicle. We therefore have no knowledge of the exact vehicle routes nor do we decide about it (see [11]).

The set of potential locations is distributed over the region of operation. Locations can have very different cost values for installation, maintenance and operation. A location can be used by a vehicle if it's close enough (e.g. does not exceed a maximal distance parameter). Vehicles and locations must be technologically compatible. A valid solution of the problem is an assignment of sequence elements to locations, such that each vehicle is supplied (remaining range never exceeds the minimal and maximal battery level). Additionally the location capacities have to be sufficient. The location capacities do not simply depend on the total number of assigned vehicles. They are derived from the maximum amount of simultaneously occupying vehicles.

3 Mixed-Integer Model Formulation

We start with the set of potential locations L. For all locations $l \in L$ we define a set K^l of discrete capacity levels. Each capacity level $k \in K^l$ has an amount of available charging slots $K_{l,k}$. The total cost value $C_{l,k}$ is derived from all fixed and variable costs over the projected time of operation (e.g. using the Net Present Value Method, see [8]).

Let $\sigma_{l,k}$ denote the decision to operate location $l \in L$ in capacity level $k \in K^l$. The planning goal is to find a cost minimal subset of locations and capacities, thus giving the following objective function:

$$\min \sum_{l \in L} \sum_{k \in K^l} C_{l,k} \cdot \sigma_{l,k} \tag{1}$$

Any location may be operated in at most one defined capacity level. $\sigma_{l,k} = 0$ for all $k \in K^l$ implies the unavailability of the location in the solution.

$$\sum_{k \in K^l} \sigma_{l,k} \leq 1 \qquad \forall l \in L \tag{2}$$

Let S be the set of all sequences. For each $s \in S$ we define a set $I^s = \{1, 2, \ldots, I_s\}$ of sequence numbers, i.e. a sequence element is identified by a tuple (s,i) with $s \in S$ and $i \in I^s$. Each sequence element consists of a set of supply options $L^{s,i}$. The locations $l \in L^{s,i}$ can be used by the vehicle, because they are close enough to the parking position and technologically compatible.

For simplification reasons we measure the vehicle energy level in remaining range (e.g. kilometers or miles). For any sequence there is a known initial remaining range X_s^0. To model the change of the energy level over time and space we use balance equations between sequence elements.

Let $q_{s,i,l}$ denote the amount of mileage collected. Remaining mileage at position $i+1$ is the previous remaining mileage $x_{s,i}$ minus the driven distance $D_{s,i}$ plus additional mileage $q_{s,i+1,l}$ at location $l \in L^{s,i}$:

$$x_{s,1} = X_s^0 + \sum_{l \in L^{s,1}} q_{s,1,l} \qquad \forall s \in S \tag{3}$$

$$x_{s,i+1} = x_{s,i} + \sum_{l \in L^{s,i}} q_{s,i+1,l} - D_{s,i} \qquad \forall s \in S,\ i \in I^s \setminus \{I_s\} \tag{4}$$

As the vehicle energy carrier is a battery, its level can never fall below zero and may also never exceed the available capacity. This is enforced by bounding the remaining mileage to the battery conditions. An even better bound is the next distance $D_{s,i}$:

$$0 \leq D_{s,i} \leq x_{s,i} \leq M_s \qquad \forall s \in S,\ i \in I^s \tag{5}$$

Let $\tau_{s,i,l}$ denote the assignment of a sequence element to a supply location (indicator variable). Each sequence element may be assigned at most once. For every sequence element (s,i) we calculate for every allowed location $l \in L^{s,i}$ the maximal possible gain of mileage $R_{s,i,l}$ based on the possible charging power, the available time and the vehicle's average power consumption. In fact, $q_{s,i,l}$ can also be less than $R_{s,i,l}$ if the remaining range $x_{s,i}$ equals the battery capacity M_s (i.e. vehicle is fully charged although there would be time left for additional range).

$$q_{s,i,l} \leq R_{s,i,l} \cdot \tau_{s,i,l} \qquad \forall s \in S,\ i \in I^s,\ l \in L^{s,i} \tag{6}$$

$$\sum_{l \in L^{s,i}} \tau_{s,i,l} \leq 1 \qquad \forall s \in S,\ i \in I^s \tag{7}$$

To model the capacity constraints we use discrete time slots. For each location $l \in L$ we define a set $t \in T$ of time intervals. The set $C^{l,t}$ contains all tuples (s,i) of sequence elements in time interval $t \in T$. In other words, if the sequence element (s,i) is associated to a location $l \in L^{s,i}$, one capacity unit is needed for this vehicle in time slot $t \in T$.

We formulate this capacity constraint as follows:

$$\sum_{(s,i)\in C^{l,t}} \tau_{s,i,l} \leq \sum_{k\in K^l} \sigma_{l,k} \cdot K_{l,k} \qquad \forall l \in L,\ t \in T \tag{8}$$

Extension As additional constraint for practical applications, we consider the maximal allowed location utilization over time. Given the duration T_t^{Slot} for any time slot $t \in T$ we restrict the average utilization for location $l \in L$ to a ratio of $0 \leq U_l^{Max} \leq 1$:

$$\frac{\sum_{t\in T}\left(T_t^{Slot} \cdot \left(\sum_{(s,i)\in C^{l,t}} \tau_{s,i,l}\right)\right)}{\sum_{t\in T} T_t^{Slot}} \leq U_l^{Max} \cdot \left(\sum_{k\in K^l} \sigma_{l,k} \cdot K_{l,k}\right) \qquad \forall l \in L \tag{9}$$

4 Complexity

To analyze the complexity of finding an optimal solution for a given model instance, we look at at the *Set Covering Problem*, which is a known NP-hard problem (see [4]). Given a zero-one matrix with m rows and n columns, we want to cover all rows by choosing a subset of all columns. The number of selected columns should be minimal, that means there is no subset of all columns with smaller size which covers all rows:

$$\min \sum_{i=1}^{n} x_i \tag{10}$$

$$s.t. \sum_{j=1}^{n} a_{i,j} x_j \geq 1 \qquad \forall i = 1,2,\ldots,m \tag{11}$$

$$x_j \in \{0,1\} \tag{12}$$

We now transform a given *Set Covering* instance into our model. We set $L = \{1,2,\ldots,n\}$, so every column is one potential location. We construct a single sequence s with one sequence element for any row ($I^s = \{1,2,\ldots,m\}$). We set $D_{s,i} = 1$ for all $i \in I^s$, so that the association to exactly one location (column) is needed for every sequence element.

The set of available locations $L^{s,i}$ contains all columns which can cover the current row ($L^{s,i} = \{1 \leq i \leq n \mid a_{i,j} = 1\}$). We consider exactly one capacity level for any location and set the costs to $C_{l,k} = 1$.

Any optimal solution to our model is also an optimal solution to the covering instance, as there is exactly one column selected for every row with the minimal number of possible locations (columns). Any solution algorithm to our model can also solve any given *Set Covering* instance, therefore our problem is NP-hard to solve.

5 Conclusions and Further Research

Our contribution is a new model approach for the cost-optimal selection of charging locations for electric vehicles. Demand is modeled as sequences of vehicle movements and parking activities. Considering the time of parking and a maximum allowed distance, vehicles can be associated to charging locations with a maximum possible increase of range. The objective function is to find a cost-minimal subset of all potential locations. We have shown the problem to be NP-hard and will concentrate on heuristic solution strategies in further research.

References

1. Luce Brotcorne, Gilbert Laporte, and Frédéric Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3): 451–463, 2003.
2. CC Chan. The state of the art of electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, 95(4): 704–718, 2007.
3. R.Z. Farahani and M. Hekmatfar. *Facility Location: Concepts, Models, Algorithms and Case Studies*. Springer Verlag, 2009.
4. Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.
5. C. Leitinger and G. Brauner. Nachhaltige Energiebereitstellung fuer elektrische Mobilitaet. *e & i Elektrotechnik und Informationstechnik*, 125(11): 387–392, 2008.
6. K. Morrow, D. Karner, and J. Francfort. Plug-in hybrid electric vehicle charging infrastructure review. *Final report Battelle Energy Alliance, US Department of Energy Vehicle Technologies Platform–Advanced Testing Activity*, 2008.
7. H. Neudorfer, A. Binder, and N. Wicker. Analyse von unterschiedlichen Fahrzyklen fuer den Einsatz von Elektrofahrzeugen. *e & i Elektrotechnik und Informationstechnik*, 123(7): 352–360, 2006.
8. S.A. Ross, R. Westerfield, and B.D. Jordan. *Fundamentals of corporate finance*. Tata McGraw-Hill, 2008.
9. D.A. Schilling, V. Jayaraman, and R. Barkhi. A review of covering problems in facility location. *Location Science*, 1(1): 25–55, 1993.
10. Jasna Tomic and Willett Kempton. Using fleets of electric-drive vehicles for grid support. *Journal of Power Sources*, 168(2): 459–468, 2007.
11. Paolo Toth and Daniele Vigo, editors. *The vehicle routing problem*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

Optimal Evacuation Solutions for Large-Scale Scenarios*

Daniel Dressler, Gunnar Flötteröd, Gregor Lämmel, Kai Nagel, and Martin Skutella

1 Introduction

Evacuation, the process of moving people out of potentially dangerous areas, is a key response to many threats. *Planning* such an evacuation is therefore important, especially in large-scale emergencies, where routing becomes non-trivial. This paper deals with the optimization and simulation of the evacuation process. We draw our data from the study of the city of Padang in Indonesia, with its high threat of tsunami waves.

Problem Definition. We consider the EVACUATION PROBLEM (EP), both in a deterministic and a stochastic setting. An instance consists of a directed graph $G = (V,A)$, flow rate capacities $u : A \to \mathbb{R}_{\geq 0}$, travel times $\tau : A \to \mathbb{R}_{\geq 0}$, and optionally time windows $W : A \to \{[i, j) : i < j\}$, restricting when an arc is available at all. Demands $d : V \to \mathbb{R}$ determine which vertices are sources and which are sinks. We assume that each sink is essentially uncapacitated. In the standard *flow over time* (or dynamic flow) model [2], the flow rate $f(a, \theta)$ enters the tail of the arc at time θ and leaves the head of the arc at time $\theta + \tau(a)$. Capacity constraints limit this flow rate. This model captures the essence of a number of people starting at certain vertices and moving towards safe vertices, while the arc transit times create

Daniel Dressler, Martin Skutella
Institute for Mathematics, TU Berlin, Str. des 17. Juni 136, 10623 Berlin, Germany, e-mail: {dressler, skutella}@math.tu-berlin.de

Gunnar Flötteröd
Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, e-mail: gunnar.floetteroed@epfl.ch

Gregor Lämmel, Kai Nagel
Institute for Land and Sea Transport Systems, TU Berlin, Salzufer 17-19, 10587 Berlin, Germany, e-mail: {laemmel, nagel}@vsp.tu-berlin.de

* Supported by the Federal Ministry for Education and Research (BMBF) under grants 03NAPI4, 03SKPAI6 ("Advest") and 03G0666E ("last mile") and by the DFG Research Center MATHEON *Mathematics for key technologies* in Berlin.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_38,

time-dependent movement and interaction. The desired output is a feasible flow over time satisfying the demands of the sources with *minimum total travel time*. The stochastic model differs from its deterministic counterpart in that flow units are assigned probabilistically to routes, but no randomness in the network parameters (free flow travel times, capacities) is accounted for.

Literature Overview. The EP can be solved either based on a deterministic model (with every flow unit being deterministically allocated to a unique path with unique characteristics) or a stochastic model (with every flow unit selecting a path from a choice set with a certain probability, and the path characteristics being stochastic as well). We first consider rigorous mathematical programming approaches that assume a deterministic model and then consider simulation based approaches that also cope with stochasticity, although merely in a heuristic manner.

Evacuation Planning with Deterministic Models. Hamacher and Tjandra [3] survey objectives in evacuation planning and discuss flow over time models in detail. A justification for the objective of the EP is given by [5], because minimizing the total travel time also maximizes the amount of flow that has already reached the sinks *at each time step* (if we ignore time windows). This property also defines an Earliest Arrival Flow (EAF). Most practical approaches to flow over time problems rely on time-expanded networks, with the notion of time built explicitly into the graph, at the price of a pseudo-polynomial size. A time-expanded network uses one copy of the original network per time step, with a copy of an arc a of length $\tau(a)$ pointing from time layer $\theta \in \mathbb{Z}$ to time layer $\theta + \tau(a)$. Thus, static flows on the time-expanded network correspond to flows over time and vice versa. Then, a standard MINIMUM COST FLOW computation on the time-expanded network, with costs equal to the transit times, yields an answer to the EP.

[12, 13] considers specialized algorithms for the EP and related problems, in particular when the network parameters may change over time. Our combinatorial algorithm uses similar ideas to exploit the repeating structure of the time-expanded network, but works on different assumptions. Furthermore, our large-scale instances require additional techniques to achieve acceptable run times. The QUICKEST TRANSSHIPMENT PROBLEM is a relaxation of the EP, which just asks for the minimum egress time. A highly non-trivial but polynomial time algorithm for this problem is given by [4]. [1] present a polynomial time approximation scheme using a coarser time-expanded network. Their approach also yields approximate EAFs. These results are certainly useful, but the approximation theoretically requires higher precision than what is needed in practice.

Evacuation Planning with Stochastic Models. The EP can also be considered from the perspective of dynamic traffic assignment [DTA; e.g., 10]. In order to account for stochasticity, we concentrate on simulation-based approaches that generate realizations of possible network states. We consider two possible directions to look into: The simulation of a Nash equilibrium (NE) and a marginal social cost based approach (MSCB) which in theory approximates the system optimal assignment.

An approximate NE can be simulated by alternately evaluating a route choice model and a network loading model until mutual consistency is attained. This procedure has a long history in simulation-based DTA [8]. Essentially, each simulated

evacuee iteratively optimizes its personal evacuation plan. After each iteration, every evacuee calculates the cost of the most recently executed plan. Based on this cost, the evacuee revises the most recently executed plan. Some evacuees generate new, "best-response" plans using a time-dependent router. If the cost function uses the travel time of the last iteration, the system converges towards an approximate NE, where the main source of imprecision is due to the stochastic environment in which the best responses are computed.

[9] show that an identical simulation logic can be applied to the simulation of a MSCB assignment if the link travel times are replaced by the marginal link travel times, which essentially represent the increase in total travel time generated by a single additional agent entering a link. [6] present a both computationally and conceptually very simple approximation of these marginal travel times.

Our Contribution. We solve the EP combinatorially using a customized MINIMUM COST FLOW algorithm. We then refine the solution in the simulation either towards an Nash Equilibrium (NE) or using the marginal social cost based approach (MSCB). This tests and accounts for the robustness of the deterministically computed plans in a stochastic environment.

2 The Instance and its Solution

We demonstrate our results on an instance that models a tsunami threatening the city of Padang in Indonesia. In case of a tsunami, the population should evacuate the shore line and flee to higher areas. We have a detailed network describing downtown Padang and trim this to the area below the height threshold. The resulting network has 4,444 nodes and 12,436 arcs, covering roughly 32 square kilometers.

It is assumed that all inhabitants leave on foot, with an average walking speed of 1.66 m/s. Derived from this, each arc has a flow rate capacity proportional to the smallest width of the corresponding street. The simulation uses a storage capacity that is computed from the area of the street. The scenario assumes a tsunami warning at 3:00 am, so that the 254,970 inhabitants all start at their respective homes. The first flooding starts at 3:20 am and water covers almost a third of the modeled area by 3:40 am. The evacuation plan completes around 3:50 am. The time windows remove arcs from the network when they are first covered by water.

Optimal Combinatorial Solution. Our algorithm for the EP has the main benefits that it can handle fine discretizations and almost arbitrarily large time horizons, and does not require an estimate of the egress time. The memory consumption is modest compared to standard algorithms. We need to store the underlying instance for the time-expanded network only once. We can also store the flow over time, i.e., the function for each arc, in intervals with constant values. This not only reduces memory consumption, the intervals also enable a different algorithmic approach:

The associated MINIMUM COST FLOW PROBLEM in the time-expanded network can be solved by the SUCCESSIVE SHORTEST PATH algorithm, which starts with a zero flow and iteratively adds flow on shortest paths to it. In our case, the shortest path computation only needs to determine the earliest reachable copy of a sink. For this, we need to propagate reachability in the network, which we also store

in intervals: If a vertex v is reachable during some interval $[\theta_1, \theta_2)$, and there is some arc $a = (v, w)$, then w will be reachable at times $[\theta_1 + \tau(a), \theta_2 + \tau(a))$. This assumes that the time-expanded copies of arc a have positive capacities for the entire interval. Otherwise, one can efficiently compute the intervals to propagate from the flow intervals.

We use a breadth-first search starting with reachable intervals $[0, \infty)$ from all (not yet empty) sources. If there are multiple sinks, we can also find multiple shortest paths, and we can hope to add flow on all of them. We also tried alternative search styles (e.g., starting at the sinks, or a simultaneous search from the sources and the sinks), but we found the forward search from the sources to deliver reliably good results. Dynamically joining consecutive intervals in the search also helps. In addition, we may guess shortest paths by repeating successful paths that arrived one time step earlier. Indeed, in the case of a single source, an optimum solution can be found by repeating a certain set of paths as long as possible. The addition of repeated paths to our algorithm closely mirrors this behavior.

Simulation-Based Approach. The simulation framework is based on the MATSim DTA microsimulation [e.g., 7, for the evacuation version]. It essentially implements the same network flow model as assumed in the mathematical programming approach, only that the integer flow units are now replaced by microscopically simulated agents with potentially complex internal behavioral rules.

When feeding the solution obtained with the combinatorial approach into MATSim, we expect the solution quality to deteriorate because of the stochastic system environment. We then refine these plans within that environment.

In particular, the EAF solution contains an exit plan for each agent, including a departure time, which also determines the order in which the agents should leave the sources.These plans become the starting solution for MATSim, except that the departure times are removed, i.e. all agents attempt to depart simultaneously. The assumption here is that controlled staggered departures could not be maintained under real-world conditions. As a result, agents following their plans might suddenly meet in a bottleneck, which would not have occurred with the original departure times.

From here, two different directions are considered inside MATSim: The simulation of a Nash equilibrium (NE), and a marginal social cost based (MSCB) approach. For the NE, we deploy the iterative technique described by [8], whereas for the MSCB, we use the approximate solution procedure of [6].

3 Results and Conclusion

We implemented our combinatorial algorithm for the EP in Sun Java 1.6 and use the network simplex implementation of ILOG CPLEX 12.1 for comparison.

We present results for the instance of Padang, as well as another instance modeling a 20-story building in high detail. It consists of 5,966 nodes and 21,937 edges, but only 369 people are inside. Our findings are summarized in Table 1. Note that in all cases we had to provide CPLEX with a time-expanded network for a specific time horizon T, which we chose near the optimum T^*.

Instance				CPLEX 12.1		our algorithm	
	step Δ	T^*	T	Time	Mem	Time	Mem
Padang	10s	313	350	0:08:30	0.8 GB	0:19:05	1.1 GB
Padang	3s	1051	1100	0:47:56	2.2 GB	1:03:13	1.1 GB
Padang	1s	3221	3225	4:32:48*	6.8 GB	2:22:09	1.2 GB
Building	1s	809	850	0:23:36	3.2 GB	0:00:20	0.4 GB

Table 1 Reported times are user times in h:mm:ss on a PC with a 2x2.13 GHz Core2Duo CPU, running openSUSE 11.1 (x86_64). (*Estimated from 3:29:51 on a 2x.3.0 GHz Core2Duo CPU.)

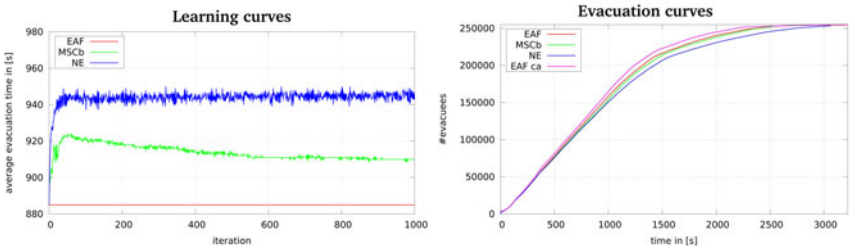
The general conclusion is that fewer agents and a greater time horizon favor our algorithm. The former effect is seen very clearly in the building instance, the latter in the Padang series. The following figures for the Padang instance with 1 second time steps might explain the good results: Despite the time horizon of 3221, in the worst case the reachability is stored in $85|V|$ intervals, the flow in $73|A|$ intervals. Finally, the many possible sinks (exits to higher ground) let the algorithm find 100 or more suitable paths at once, greatly decreasing the number of iterations required.

We then used the routes computed combinatorially for Padang ($\Delta = 1s$) as initial routes in MATSim. The computed average travel time was $844s$, while the same routes in MATSim result in $885s$. This difference of 5% seems quite acceptable, and may stem from the rounded capacities and discarded starting times, leading to additional disturbances.

The iterative process of MATSim then modifies this solution, reaching almost stable states with avg. travel times $945s$ (NE) and $909s$ (MSCB), respectively. The learning curves for NE and MSCB, leading to these numbers, are shown in Fig. 1(a). For comparison, the results for the EAF solution in MATSim are also plotted over the 1000 learning iterations. The MSCB solution is worse than the EAF solution because the MSCB approach incorporates several simplifying assumptions. Meanwhile, it is plausible that the NE approach leads to a solution that is worse than that from the MSCB approach: This is the well-known price of anarchy [11].

Fig. 1(b) shows the evacuation curves, i.e. the cumulative number of evacuated persons as a function of time. Despite the difference in the numbers above, these curves look astonishingly similar. The curve denoted by "EAF ca" corresponds to the theoretical arrival curve calculated by the combinatorial algorithm. This may indicate that, despite the difference in approaches, we reach structurally similar solutions.

Concluding, we compute solutions that are optimal in the analytical model, which differs from the model of the learning framework in some details. We observe that the EP solution changes only slightly when replayed without the departure times of the individual flow units in the stochastic simulation environment. This leads us to believe that it is indeed a good solution. The NE and MSCB solutions of the simulation framework are also of good quality, which is evident from the optimum combinatorial solution.



(a) Avg. evacuation time over iteration number (b) Arrival curves

Fig. 1 Results of the MATSim simulation for EAF, NE and MSCB.

Let us remark that NE and MSCB solutions of similar quality[2] can also be obtained by the simulation-framework without optimized initial routes.

References

1. Lisa Fleischer and Martin Skutella. Quickest flows over time. *SIAM Journal on Computing*, 36: 1600–1630, 2007.
2. L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.
3. H.W. Hamacher and S.A. Tjandra. Mathematical modelling of evacuation problems – a state of the art. In *Pedestrian and Evacuation Dynamics*, pages 227–266, 2002.
4. B. Hoppe and É. Tardos. The quickest transshipment problem. *Mathematics of Operations Research*, 25: 36–62, 2000.
5. J. Jarvis and H. Ratliff. Some equivalent objectives for dynamic network flow problems. *Management Science*, 28: 106–108, 1982.
6. G. Lämmel and G. Flötteröd. Towards system optimum: finding optimal routing strategies in time dependent networks for large-scale evacuation problems. Volume 5803 of *LNCS (LNAI)*, pages 532–539, Berlin Heidelberg, 2009. Springer.
7. G. Lämmel, D. Grether, and K. Nagel. The representation and implementation of time-dependent inundation in large-scale microscopic evacuation simulations. *Transportation Research Part C: Emerging Technologies*, 18: 84–98, February 2010.
8. K. Nagel and G. Flötteröd. Agent-based traffic assignment: going from trips to behavioral travelers. In *Proceedings of 12th International Conference on Travel Behaviour Research*, Jaipur, India, December 2009. Invited resource paper.
9. S. Peeta and H.S. Mahmassani. System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Annals of Operations Research*, 60: 81–113, 1995.
10. S. Peeta and A.K. Ziliaskopoulos. Foundations of dynamic traffic assignment: The past, the present and the future. *Networks and Spatial Economics*, 1: 233–265, 2001.
11. T. Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, May 2005.
12. S.A. Tjandra. Earliest arrival flow with time dependent capacity approach to the evacuation problems. Technical report, TU Kaiserslautern, 2001.
13. S.A. Tjandra. *Dynamic Network Optimization with Application to the Evacuation Problem*. PhD thesis, TU Kaiserslautern, 2003.

[2] Avg. travel times if started with shortest path as initial solution converges to 943*s* for NE and 917*s* for MSCB, respectively.

An Integrated Vehicle-Crew-Roster Problem with Days-Off Pattern

Marta Mesquita, Margarida Moz, Ana Paias, and Margarida Pato

Abstract The integrated vehicle-crew-roster problem with days-off pattern aims to simultaneously determine minimum cost sets of vehicle and daily crew schedules that cover all timetabled trips and a minimum cost roster covering all daily crew duties according to a pre-defined days-off pattern. This problem is modeled as a mixed binary linear programming problem. A heuristic approach with embedded column generation and branch-and-bound techniques within a Benders decomposition is proposed. The new methodology was tested on real instances and the computational results are promising.

1 Introduction

The integrated vehicle-crew-roster problem aims to simultaneously assign drivers of a company to vehicles and vehicles to a set of pre-defined timetabled trips that cover passenger transport demand in a specific area, during a planning horizon. Due to the complexity of the corresponding combinatorial optimization problem, it is usually tackled on a sequential basis. Given a set of timetabled trips, the vehicle scheduling produces a set of daily schedules for the vehicles that perform all trips. The crew scheduling defines the daily crew duties covering the vehicle schedules while respecting minimum and maximum spread, maximum working time without a

Marta Mesquita
ISA-UTL / CIO, Tapada da Ajuda, 1349-017 Lisboa, Portugal, e-mail: marta@math.isa.utl.pt

Margarida Moz
ISEG-UTL / CIO, Rua do Quelhas 6, 1200-781 Lisboa, Portugal e-mail: mmoz@iseg.utl.pt

Ana Paias
DEIO-FCUL / CIO, BLOCO C6, Piso 4, Cidade Universitária, 1749-016 Lisboa, Portugal e-mail: ampaias@fc.ul.pt

Margarida Pato
ISEG-UTL / CIO, Rua do Quelhas 6, 1200-781 Lisboa, Portugal e-mail: mpato@iseg.utl.pt

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_39,

break, minimum and maximum break duration, maximum number of changeovers, etc. Finally, for the planning horizon, crew duties are assigned to the company's drivers leading to a roster that must comply with labor and institutional norms.

The problem addressed in this paper is the integrated vehicle-crew-roster problem with days-off pattern (VCRP). The VCRP solution consists of daily vehicle and crew schedules, covering all timetabled trips, and a roster, following a cyclic days-off pattern in use at a mass transit company for a specific group of drivers.

For a time horizon of 7 weeks (49 days), this days-off pattern includes 4 rest periods of 2 consecutive days-off and 2 rest periods of 3 consecutive days-off. These two rest periods contain a Saturday and occur in sequence with 5 workdays in between. The remaining work periods have 6 consecutive workdays. Table 1 displays such a pattern through a $0-1$ matrix, where 0 stands for day-off and 1 for workday. Each row of the matrix corresponds to a weekday. Seven consecutive columns correspond to 7 weeks of the time horizon, being the last day in column i $(i = 1, ..., 6)$ followed by the first day in column $i+1$ and the last day in column 7 followed by the first day of column 1. A 49 days schedule may begin in row 1 of any column.

Table 1 Cyclic days-off pattern

	1	2	3	4	5	6	7
Monday	0	1	1	1	1	1	0
Tuesday	0	0	1	1	1	1	1
Wednesday	1	0	0	1	1	1	1
Thursday	1	1	0	0	1	1	1
Friday	1	1	1	0	0	1	1
Saturday	1	1	1	1	0	0	1
Sunday	1	1	1	1	0	0	1

The main objective of the VCRP derives from the minimization of vehicle and driver costs. Since ŞunpopularŤ duties and overtime are undesirable they should also be minimized and equitably distributed among the drivers. This paper proposes an heuristic approach based on Benders decomposition that iterates between the solution of an integrated vehicle-crew scheduling problem and the solution of a rostering problem. To the authors' knowledge, no integration methodologies have been proposed on this context. Although for the integration of airline operations, Benders decomposition methods have already been applied by [1], [2] and [4].

A mathematical formulation for the VCRP is presented in section 2. Section 3 describes the solution approach. Section 4 shows computational results and section 4 presents some conclusions.

2 Mathematical Formulation

During a planning horizon H, partitioned into 49 days, a set M of drivers must be assigned to a fleet of vehicles from $|D|$ depots in order to perform a set of timetabled trips (trips for short). For each trip the starting and ending times and locations are known. Let N^h be the set of trips to be performed on day h. Trips i and j are compatible if the same vehicle can perform both in sequence. Between i and j a deadhead trip may occur where the vehicle runs without passengers. Let I^h be the set of compatible pairs of trips, on day h. Decision variables, related to the scheduling of vehicles, are defined as $z_{ij}^{dh} = 1$ if a vehicle from depot d performs trips i and j in sequence on day h, or 0 otherwise. Depots may be seen as special trips so as z_{dj}^{dh} and z_{id}^{dh} represent, respectively, the deadhead trip from depot d to trip j and the deadhead trip from i to depot d. Vehicle costs, c_{ij}, are related with fuel consumption and vehicle maintenance. Daily vehicle schedules must be covered with daily crew duties. Let L^h be the set of crew duties of day h and $L_{ij}^h \subseteq L^h$ the set of crew duties covering the deadhead trip from the end location of trip i to the start location of j and covering trip j, on day h. Let variable $w_\ell^h = 1$ if crew duty ℓ is selected on day h, or 0 otherwise. Each crew duty has to be assigned to a driver which works according to one of the seven pre-defined days-off schedules. Let $x_s^m = 1$ if driver m is assigned to schedule s, or 0 otherwise, and let $y_\ell^{mh} = 1$ if driver m performs crew duty ℓ on day h, or 0 otherwise. Management often wants to know the minimum workforce required to operate the fleet of vehicles, so as to assign drivers to other departments of the company or to replace those absent. Such policy results in minimizing crew duty costs e_ℓ associated to variables $w_\ell^h = 1$ and costs r^m associated to $x_s^m = 1$.

To satisfy labor rules crew duties are grouped into early crew duties, L_E^h, starting before 3:30 p.m. and late crew duties, L_A^h, starting after 3:30 p.m. The set of crew duties is also partitioned into short duties, L_T^h, which have a maximum spread of 5 hours (without a break), normal duties with spread $\in [5,9]$ hours and long duties, L_O^h, with spread $\in [9,10.25]$ hours (with overtime). To balance workload, penalties λ_T and λ_O are associated with variables δ_T and δ_O that represent, respectively, the maximum number of short and long duties assigned to a driver.

The VCRP can be formulated through the following MILP model:

$$min \sum_{h\in H}\Big(\sum_{d\in D}\sum_{(i,j)\in I^h} c_{ij}z_{ij}^{dh} + \sum_{\ell\in L^h} e_\ell w_\ell^h\Big) + \sum_{m\in M}\sum_{s\in S} r^m x_s^m + \lambda_T\delta_T + \lambda_O\delta_O \quad (1)$$

$$\sum_{d\in D}\sum_{i:(i,j)\in I^h} z_{ij}^{dh} = 1,\ j \in N^h, h \in H \quad (2)$$

$$\sum_{i:(i,j)\in I^h} z_{ij}^{dh} - \sum_{i:(j,i)\in I^h} z_{ji}^{dh} = 0,\ j \in N^h, d \in D, h \in H \quad (3)$$

$$\sum_{i\in N^h} z_{d,i}^{dh} \le v_d,\ \forall d \in D, h \in H \quad (4)$$

$$\sum_{\ell \in L_{ij}^h} w_\ell^h - \sum_{d \in D} z_{ij}^{dh} \geq 0,\ (i,j) \in I^h, h \in H \tag{5}$$

$$\sum_{m \in M} y_\ell^{mh} - w_\ell^h = 0,\ \ell \in L^h, h \in H \tag{6}$$

$$\sum_{s \in S} x_s^m \leq 1,\ m \in M \tag{7}$$

$$\sum_{\ell \in L^h} y_\ell^{mh} - \sum_{s \in S} a_s^h x_s^m \leq 0,\ m \in M, h \in H \tag{8}$$

$$\sum_{\ell \in L_f^h} y_\ell^{mh} + \sum_{\ell \in L_g^{(h-1)}} y_\ell^{m(h-1)} \leq 1,\ m \in M, h \in H - \{1\}, f, g \in \{E, A\}, f \neq g \tag{9}$$

$$\sum_{h \in H} \sum_{\ell \in L_t^h} y_\ell^{mh} - \delta_t \leq 0,\ m \in M, t \in \{T, O\} \tag{10}$$

$$z_{ij}^{dh} \in \{0,1\},\ (i,j) \in I^h, d \in D, h \in H \tag{11}$$

$$w_\ell^h \in \{0,1\},\ \ell \in L^h, h \in H \tag{12}$$

$$y_\ell^{mh} \in \{0,1\},\ \ell \in L^h, m \in M, h \in H \tag{13}$$

$$x_s^m \in \{0,1\},\ s \in S, m \in M \tag{14}$$

$$\delta_T, \delta_O \geq 0. \tag{15}$$

The objective function measures the quality of the solution. Constraints (2) together with (3) ensure that each timetabled trip is performed, exactly once, by a vehicle that returns to the source depot. Constraints (4) are depot capacity constraints where v_d is the number of vehicles available at depot d. Constraints (5) ensure that vehicle schedules are covered by, at least, one crew. Equalities (6) impose that each crew duty, in a solution, must be assigned to one driver. Constraints (7) state that each driver is assigned to one of the seven days-off schedules or is available for other service in the company. Constraints (8), where $a_s^h = 1$ if h is a workday on schedule s, impose coherence between the assignment of a crew duty to a driver and the schedule assigned to this driver. Inequalities (9) forbid the sequence late/early duties to ensure that drivers rest a given minimum number of hours between consecutive duties. Furthermore, (9) impose a day-off period between changes of duty types. Inequalities (10) calculate the maximum number of short/long duties per driver. A decomposition approach is suggested by the three combinatorial structures included in this mathematical formulation. A multi-commodity network flow problem is related with the daily schedule of vehicles. A set covering structure defines the crews that daily cover the vehicle schedules and an assignment problem, with additional constraints, defines a roster that covers all daily crew duties.

3 Decomposition Approach

Two different temporal scheduling problems may be identified within the combinatorial structures included in the above mathematical formulation: daily integrated vehicle-crew scheduling problems and a rostering problem for the whole planning horizon H. These problems share a set of variables within constraint set (6). By taking into account the variables involved in this constraint set, a decomposition approach, based on Benders method, is proposed. The approach alternates between the solution of a master problem involving the z, w variables, a vehicle-crew scheduling problem for each day of H, and the solution of the corresponding sub-problem involving the y, x, δ variables, a rostering problem. In each iteration, the sub-problem is obtained by fixing the z and w variables at values $\overline{z}$ and $\overline{w}$ given by the optimal solution of the corresponding master problem. The dual of the linear relaxation of the resulting sub-problem is solved and a Benders cut is obtained with the corresponding optimal dual variables. In the traditional Benders algorithm this cut is added, as a new constraint, to the master problem in the next iteration. However, this new constraint involves variables related with all the days of the planning horizon and the resulting combinatorial optimization master problem is as difficult to solve as the original VCRP. Consequently, a non-exact approach is proposed where, in each iteration, the Benders cut is relaxed into the master objective function thus penalizing crew duty costs e_ℓ associated to variables $w_\ell^h = 1$. The relaxed master problem can be decomposed into $|H|$ independent integrated vehicle-crew scheduling problems. Since crew duty costs e_ℓ do not depend on the day h, these $|H|$ problems are reduced to one for each day type (1 day type for workdays and 1 or 2 for weekends). Integrated vehicle-crew scheduling problems are solved by the algorithm proposed in [3] which combines a heuristic column generation procedure with a branch-and-bound scheme. Exact standard algorithms are used to solve the linear relaxation of the rostering sub-problem. Whenever the resulting solution is not integer, branch-and-bound techniques are applied to obtain a feasible roster that along with the solution of the relaxed master correspond to a feasible solution for the VRCP.

4 Computational Experience

The test instances were derived from a bus service operating in the city of Lisbon and involve problems with 122, 168, 224, 226 and 238 trips and 4 depots. The algorithms were coded in C, using VStudio 6.0/C++. Linear programming relaxations and branch-and-bound schemes were tackled with CPLEX 11.0.

In Table 2, for each test instance, the first row corresponds to the solution obtained in the first iteration and the second row describes the best solution obtained within 10 iterations. Columns 'vehicles' and 'crews' describe the master problem solution. This columns show, respectively, the number of daily vehicles and the number of daily crew duties, short/normal/long, from Monday through Friday and in brackets for the weekend. For the corresponding sub-problem solutions, involv-

ing 7 weeks, column 'drivers' shows the number of drivers and column 'short/long' shows the maximum number of short/long duties assigned to a driver. The last two columns report on computational times, in seconds, for the first iteration and, in the best solution row's, for the 10 iterations.

Table 2 Results from 10 iterations of the decomposition algorithm (PC Pentium IV 3.2 GHz)

trips	iteration	vehicles	crews	drivers	short/long	master time(sec)	sub-prob. time(sec)
122	first	9 (6)	1/8/8 (4/3/2)	25	4/13	8	135
	best it-6	9 (6)	0/9/8 (3/5/1)	25	2/12	49	1102
168	first	17 (10)	4/14/20 (7/6/6)	56	5/14	9	746
	best it-3	17 (10)	4/17/17 (8/7/4)	54	5/13	75	5550
224	first	18 (10)	0/9/30 (1/4/11)	55	1/22	525	403
	best it-1	18 (10)	0/9/30 (1/4/11)	55	1/22	1410	2107
226	first	16 (7)	0/8/26 (0/6/9)	50	0/21	369	179
	best it-4	16 (7)	0/10/24 (0/7/8)	48	0/20	2232	382
238	first	22 (11)	5/13/36 (11/8/8)	77	5/18	175	9967
	best it-7	22 (11)	5/12/37 (10/7/10)	76	5/19	604	19155

One can see that, except for problem 224, the decomposition approach improves the first solution which corresponds to a sequential approach. In fact, by introducing Benders cuts in the objective function, crew duties are adjusted in the master problem, thus inducing better sub-problem solutions in what concerns the number of drivers and/or the number of short and long duties assigned to a driver.

5 Conclusions

The proposed decomposition approach for the VCRP outperforms the traditional sequential approach. The feedback given by Benders cuts guided the building of the daily vehicle-crew schedules thus leading to balanced workload rosters with fewer drivers.

References

1. J-F. Cordeau, G. Stojković, F. Soumis, and J. Desrosiers. Benders decomposition for simultaneous aircraft routing and crew scheduling. *Transportation Science*, 35: 375–388, 2001.
2. A. Mercier and F. Soumis. An integrated aircraft routing, crew scheduling and flight retiming model. *Computers & Operations Research*, 34: 2251–2265, 2007.
3. M. Mesquita and A. Paias. Set partitioning/covering-based approaches for the integrated vehicle and crew scheduling problem. *Computers & Operations Research*, 35(4): 1562–1575, 2008.
4. N. Papadakos. Integrated airline scheduling. *Computers & Operations Research*, 36: 176–195, 2009.

Improved Destination Call Elevator Control Algorithms for Up Peak Traffic*

Benjamin Hiller, Torsten Klug, and Andreas Tuchscherer

Abstract We consider elevator control algorithms for destination call systems, where passengers specify the destination floor when calling an elevator. In particular, we propose a number of extensions to our algorithm BI [4] aimed to improve the performance for high intensity up peak traffic, which is the most demanding traffic situation. We provide simulation results comparing the performance achieved with these extensions and find that one of them performs very well under high intensity traffic and still satisfactorily at lower traffic intensities.

1 Introduction

The task of an elevator control algorithm is to schedule a group of elevators to serve the passenger flow efficiently and offer a good service level. In office buildings, the critical traffic situation [2] is high intensity morning *up peak traffic*, where all (or at least most) of the passengers enter the building at the entrance floor and want to travel to the upper floors. Most elevator installations employ a *2-button* or *conventional* control system, in which a passenger tells the elevator control his desired travel direction only. Incoming passengers will board the elevators that are available at the entrance floor. Once the passengers boarded and entered their destinations, all the elevator control can do is to visit these floors in order and return to the entrance floor. It is well known that an elevator has to stop at a large fraction of the upper floors [2], leading to a long roundtrip time, yielding in turn long waiting times since the entrance floor is visited less frequently.

To improve the performance of elevator systems, *destination call systems* were introduced in the last 20 years. In such a system, a passenger registers his destina-

Zuse Institute Berlin, Takustraße 7, D–14195 Berlin, Germany
e-mail: {hiller,klug,tuchscherer}@zib.de

* Supported by the DFG Research Center MATHEON *Mathematics for key technologies*.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_40,

tion floor right at his start floor instead of the travel direction only. In response, the elevator control assigns the passenger to a serving elevator.[2] Destination call systems offer two main advantages: First, the uncertainty about the destination floors is removed, so the control algorithm has a more complete picture of the current situation which may help to find better control decisions. Second, it offers more control possibilities since it assigns passengers to elevators. In up peak traffic, this may be used to group passengers by destination floors, reducing the number of stops on a roundtrip for each elevator.

Related Work. Most of the literature on elevator control deals with conventional systems, see e. g., Barney [2] for an overview. Although Closs [2] proposed destination call systems in 1970, they were first implemented by Schindler [6] in 1990, followed by other manufacturers. In [4, 3], we described our heuristic BI, a variant of which has been implemented by our industry partner Kollmorgen Steuerungstechnik GmbH. As far as we can tell from the publications, all of these algorithms are cost-based reoptimization algorithms with no provisions to reduce roundtrip times. So far, the only known method to group by destination floor is *zoning*, where each elevator serves only a subset of destination floors that has been determined beforehand.

Contribution. We propose a number of extensions to our BI algorithm that exploit the control possibilities of destination call systems to obtain short roundtrip times and thus to improve the performance of the elevator group for up peak traffic. In contrast to zoning, our extensions determine dynamically which elevator serves which destination floors. It turns out that it is possible to achieve good high intensity performance *and* good low intensity performance with the same algorithm.

2 Control Algorithms for Destination Call Systems

The algorithms we study are based on reoptimization. At each reoptimization instant, a *snapshot problem* capturing the current system state is solved heuristically. Each snapshot problem is described by the state of each elevator, the set of floors each elevator needs to stop at due to loaded passengers, and a set of not yet served (destination) calls. The task of an elevator control algorithm is to distribute the unserved calls among the elevators and to compute for each elevator a schedule serving the assigned calls and visiting the destination floors for the loaded passengers. The resulting schedules should minimize some cost function described below.

When an elevator visits a floor, it needs to serve all destination calls waiting there that have been assigned to it and that travel in its leaving direction. We therefore collect all calls at a floor assigned to the same elevator and traveling in the same direction in an *assigned request*, that may only be served by the respective elevator. In addition, each yet unassigned destination call (i. e., one that has been registered

[2] It is also conceivable that the elevator control signals which destination floors an arriving elevator is going to serve shortly before its arrival [3]. Due to lack of space, we omit our results for this kind of system.

after the last reoptimization) constitutes an *unassigned request* that may be served by any elevator.

A *feasible schedule* is a sequence of stops at floors, specifying at each stop which requests are picked up. The set of picked up requests has to include the matching assigned request as described above and may include additional unassigned requests. Picking up a request means that the corresponding calls are now loaded and that the elevator has to stop at their destination floors. The sequence of stops has to be such that the elevator does not change its travel direction as long as passengers are loaded. A set of feasible schedules constitutes a *dispatch*, if every request is served by exactly one elevator.

Each schedule is evaluated using a cost function including the waiting and travel time for each call. The waiting time and the travel time of a passenger is the time span between the release of the call and the arrival of the serving elevator at the start floor and destination floor, respectively. Both can be estimated using parameters of the elevator system. Moreover, we assume that boarding and leaving takes a constant time for each passenger. The overall cost function is a weighted sum of the waiting and travel times plus a penalty for each call that is loaded beyond cabin capacity.

Our investigations are based on our algorithm BI, which is a cost-based best-insertion procedure. In order to compute a dispatch for a snapshot problem, it tries all feasible insertions of *unassigned requests* into the existing schedule of each elevator that already serves all assigned requests, choosing the elevator and insertion position with least additional cost.

3 Techniques for Limiting the Round Trip Time

We already argued in the introduction that a short round trip time, henceforth abbreviated RTT, is essential for good up peak performance. It turns out that the reoptimization w. r. t. to the above cost function alone does not lead to short round trip times under high intensity traffic. The reason for this behavior is that in each snapshot it is more expensive to delay service of a request by assigning it to an elevator arriving later than to serve it with the elevator arriving next, but with an additional stop. In a sense, this is due to the fact that the impact of an assignment decision on *future requests*, which are suffering from long roundtrip times, is not taken into account.

Preliminary experiments showed that penalizing long RTTs in the cost function does not improve RTTs for high intensity traffic. Instead of penalizing long RTTs, we may also avoid them by restricting the assignments of requests to elevators such that each elevator does only few stops. To this end, we modify BI to consider for a request r starting on the entrance floor only a subset $A(r)$ of *admissible* elevators for insertion. In the following, we say that an elevator *serves a floor* in a given schedule, if it picks up a request going to this floor at its stop at the entrance floor.

Our first idea for obtaining a suitable subset $A(r)$ is to distribute the destination floors evenly among the elevators. To achieve this, we can use the following rule:

If the destination floor of r is already served by one of the elevators, say elevator e, then $A(r) = e$. Otherwise, $A(r)$ is the set of elevators currently serving the least number of destination floors. We will call this rule the *GF rule*, since its main idea is essentially that of the algorithm GREEDYFIT for the online bin coloring problem [4, 3]. Observe that the GF rule does not take into account limited cabin capacity.

It is to be expected that the GF rule performs badly for low intensity traffic, since it will first assign each elevator one destination floor before one elevator is allowed to serve two floors. To overcome this issue, we propose another way of limiting the RTT we call the *destination floor limit* or DFL for short. The idea is to allow each elevator to serve at most k destination floors with $k \geq \lceil |F|/|E| \rceil$ where F is the set of floors above the entrance floors and E is the set of elevators in the group. Note that for implementing this idea it is not sufficient to make BI consider only schedules obeying the limit k. The reason is that it may happen that all elevators serve not more than k destination floors each, but still do not cover all destination floors. We can avoid these situations by suitably choosing $A(r)$. Let $F(e)$ be the set of destination floors already served by elevator e and F_s be the set of destination floors served by any elevator plus the destination floor of request r. We think of each elevator as having k *slots*, each of which can be used for one destination floor. The total number of unused slots is then given by $\kappa := \sum_{e \in E} \big(k - |F(e)|\big)$. The bad situation sketched above cannot happen if after assigning request r we have at least as many free slots as unserved destination floors, the number of which is given by $|F \setminus F_s|$. When the DFL is used, we thus set $A(r)$ according to the following cases. If $\kappa > |F \setminus F_s|$, we let $A(r)$ be the set of elevators that either have free slots or do already serve the destination floor of r. If $\kappa = |F \setminus F_s|$, $A(r)$ is the set of elevators already serving the destination floor of r. In our simulations, we choose $k = \lceil |F|/|E| \rceil$.

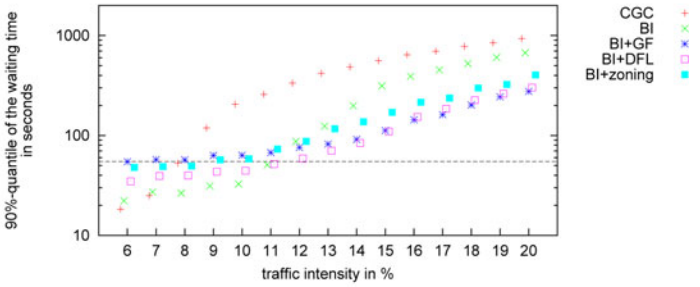
4 Computational Results

To evaluate the performance of our algorithms, we consider *real up peak* traffic, consisting to 90% of calls starting at the entrance floor, 5% calls going to the entrance floor, and 5% calls between the upper floors; it is more realistic than pure up peak traffic. We simulate 15 minutes of real up peak traffic, with intensities reaching from 6% to 20% of the building population arriving in five minutes. As benchmark values, we look at the 0.9-quantile of the waiting time experienced by calls arriving in the middle five minutes. We use the 0.9-quantile since it captures the tail of the waiting time distribution and it is more robust than the maximum. It is also used in elevator industry: A 0.9-quantile of at most 55 seconds is considered to be "fair" performance [2]. For our simulation results, we consider two example buildings / elevator systems whose data is given in Figure 1(a).

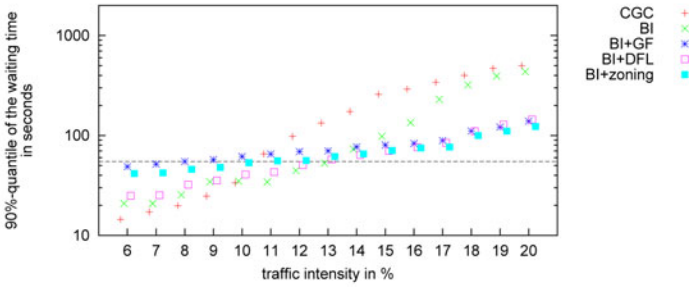
In our tests we used the following settings for the cost structure of the snapshot problem: The waiting time cost coefficient is 2 and the one for travel time is 1, reflecting that waiting time is more important than travel time. The capacity penalty corresponds to the waiting time cost of 15 minutes. In order to compare

	building A	building B
population	3300	1400
floors	12	23
elevators	8	6
cabin capacity	19	13
acceleration [m/s^2]	1.0	0.9
maximum speed [m/s]	2.0	5.0
deceleration [m/s^2]	1.0	0.9

(a) Details of the elevator systems in the buildings considered.



(b) Simulation results for building A.



(c) Simulation results for building B.

Fig. 1 Building data and simulation results.

to conventional systems, we implemented the CGC algorithm that features a good performance in most traffic situations [2]. As discussed before, for up peak traffic there is not much a control algorithm can decide so the performance of CGC should be very similar to that of other algorithms for conventional systems.

Figures 1(b) and 1(c) present our results. The algorithm CGC represents conventional systems, BI is the destination call algorithm without additional techniques; in the settings BI+GF and BI+DFL the GF rule and the DFL, respectively, have been enabled additionally.

Finally, the setting BI+zoning uses zoning to assign calls from the entrance floor to elevators. Following ideas in [7], for each building we empirically distributed upper floors among the elevators.

From the figures it is evident that CGC handles only low traffic intensities reasonably. We also see that at some point, the performance of the default BI algorithm becomes almost as bad, whereas the BI variants employing grouping techniques perform much better. As expected, BI+GF is rather bad for low traffic intensities, but very good on high intensity traffic. The same is true for BI+zoning, which performs worse than BI+GF on building A, but better than BI+GF on building B. Although BI+DFL never gives the best results, it provides good performance for all traffic intensities.

Since the algorithms that are good for low traffic intensities are rather bad for high traffic intensities and vice versa, it is probably necessary to switch the strategy depending on the load to achieve better performance for all load situations.

References

1. Gina Carol Barney. *Elevator Traffic Handbook: Theory and Practice*. Taylor and Francis, 2002.
2. Gordon David Closs. *The computer control of passenger traffic in large lift systems*. PhD thesis, Victoria University of Manchester, 1970.
3. Benjamin Hiller. *Online Optimization: Probabilistic Analysis and Algorithm Engineering*. PhD thesis, TU Berlin, 2009.
4. Benjamin Hiller and Andreas Tuchscherer. Real-time destination-call elevator group control on embedded microcontrollers. In *Operations Research Proceedings 2007*, pages 357–362. Springer, 2008.
5. Sven Oliver Krumke, Willem E. de Paepe, Leen Stougie, and Jörg Rambau. Bincoloring. *Theoret. Comput. Sci.*, 407(1–3): 231–241, 2008.
6. Jordis Schröder. Advanced dispatching: Destination hall calls + instant car-to-call assignments: M10. *Elevator World*, pages 40–46, March 1990.
7. Janne Sorsa, Henri Hakonen, and Marja-Liisa Siikonen. Elevator selection with destination control system. *Elevator World*, pages 148–155, 2006.

A Two Stage Approach for Balancing a Periodic Long-Haul Transportation Network

Anne Meyer, Andreas Cardeneo, and Kai Furmans

Abstract This paper presents a concept and planning methods to adapt lean production principles – especially regularity and balancing – to a long-haul transportation network. The two stage approach is discussed and solution methods for both stages are outlined. Summarised results of a simulation study on real network data are given to show the logistical impact of the concept.

1 Introduction

More and more industrial companies successfully bring their production in line with lean principles of the Toyota Production System. Usually, the implementation of the concept does not cover long-haul transports in the geographically widespread networks as typically found in Europe. However, high delivery frequencies (along with small transport volumes) and a high reliability in the logistics network are an indispensable condition to effectively run a lean production, see [6]. Within the research project *LogoTakt*[1] we developed an innovative logistics concept and the corresponding planning tools to address exactly these issues: High delivery frequencies can only be achieved by collaboration (aiming at the consolidation of small shipments) and the use of railway for the – in Europe usually long – main leg relations. To bring together both, the network needs to be served in a regular way, i.e. it is best to provide fixed, weekly recurring, robust transportation schedules for a defined planning horizon. Selinger et. al. summarise in [5] the goal of the project as the combination of "the positive characteristics of modern parcel transport systems, company-organized milk runs, and traditional mixed cargo systems."

Anne Meyer, Andreas Cardeneo, Kai Furmans
FZI Forschungszentrum Informatik, Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, e-mail: {meyer, cardeneo, furmans}@fzi.de

[1] This projct has partly been funded by the BMWi (BMWi, German Federal Ministry of Economics and Technology, contract number 19 G 7036 E, www.logotakt.de).

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_41, © Springer-Verlag Berlin Heidelberg, 2011

In this paper, we focus on the key innovation in the logistical concept of *LogoTakt*. We present a two stage approach which allows us to build a balanced and weekly recurring transportation plan adopting key lean principles to a long haul logistics network: regularity and standardisation, levelling of resource use, and transparency. On the first stage – on the production site level – clients choose their optimal weekly transport frequencies for different relations. To take this decision, transport volumes per relation have to be forecasted and a trade-off between holding and transportation costs has to be found. A typical dynamic lot-sizing problem, with special constraints to evaluate weekly frequencies, has to be solved. In a second step – on the network level – the network operator maps the frequency bookings of all clients to weekly transportation patterns (such as Mon-Wed-Fri for a frequency request of 3) considering the objective of levelling the volumes for the pick-up and delivery regions over the days of the week. The shift of this decision to the network level is used to avoid peaks over the course of the week and, thus, to minimise expensive stand-by capacity for pick-up and delivery transports.

A good overview of the state of the art models and solution methods for dynamic lot-sizing problems is given in [2]. The second stage problem, considered isolatedly, can be seen as a tactical network planning problem, where network services are scheduled in a balanced way (see for example [3]). Along with the subsequently planned routing, our two stage approach tackles a problem which has a certain similarity to Inventory Routing and Period Vehicle Routing Problems. Recent work in these fields also had a focus on balancing workloads (see for example [1] and [4]), however, neither of the problem types covers this issue in multi-stage transport networks.

In the remainder of this article, we introduce MIP formulations for both stages and outline the corresponding solution methods (Section 2). In Section 3 we present results of the application to a real-world supply network and we conclude with remarks about further research directions (Section 4).

2 A Two Stage Approach

As stated before, the problem on the first stage is – basically – a dynamic lot-sizing problem (see [2]) separately solved for all transport relations of a client:

Variables:	$Q_{i,t}$	lot size of product i in time step t (with $i \in \{1,\dots,I\}$ and $t \in \{1,\dots,T\}$)
	$y_{i,t}$	stock for product i in time step t
	p_m	1 if pattern m is chosen, 0 else
	$z_{pc,t}$	1 if number of pallets pc is chosen in timestep t, 0 else
	q_t, q_n	1 if lot size in t and n respectively not 0 (with $n \in \{1,\dots,T\}$), else 0
Parameters:	$PAT_{m,t}$	week pattern m in time step t, e.g. $\{1,0,1,0,1,1,0,1,0,1,...\}$ for Mon-Wed-Fri, (with $m \in M_{f_{i,j}}$: set of reasonable patterns for frequency $f_{i,j}$)
	$TP_{dc,pc}$	transport price for distance class dc and number of pallets pc
	$d_{i,t}$	demand of product i in time step t
	K^{fix}, k_i^{var}	costs per shipment and costs per unit on stock and time step t

$$\min \left(K^{fix} \cdot \sum_m \sum_t (PAT_{t,m} \cdot p_m) + \sum_t (z_{pc,t} \cdot TP_{dc,pc}) + \sum_i \sum_t (k_i^{var} \cdot y_{i,t}) \right) \quad (1)$$

$$s.t. \quad y_{i,t-1} + Q_{i,t} - y_{i,t} = d_{i,t} \qquad \forall \quad i,t|t>0 \quad (2)$$

$$\sum_i Q_{i,t} - bigM \cdot (\sum_m PAT_{m,t} \cdot p_m) \leqq 0 \qquad \forall \quad t \quad (3)$$

$$\sum_m p_m = 1 \quad (4)$$

$$\sum_{pc} pc \cdot z_{pc,t} = \sum_i Q_{i,t} \qquad \forall \quad t \quad (5)$$

$$\sum_{pc} z_{pc} = 1 \qquad \forall \quad t \quad (6)$$

$$\sum_i Q_{i,n} \leqq \sum_i Q_{i,t} + bigM \cdot (1 - q_t) \qquad \forall \quad t,n|t<n \quad (7)$$

$$\sum_i Q_{i,n} \geqq \sum_i Q_{i,t} + bigM \cdot (1 - q_n) \qquad \forall \quad t,n|t<n \quad (8)$$

$$q_t = \sum_m PAT_{m,t} \cdot p_m \qquad \forall \quad t \quad (9)$$

$$y_{i,0} \leqq \text{init stock}, \qquad y_{i,T} \leqq \text{final stock} \qquad \forall \quad i \quad (10)$$

$$z, pc, q \in \{0,1\}, \qquad y, Q \geqq 0 \quad (11)$$

The objective function (1) minimises the total sum of transportation and holding costs, the block of constraints (2) contains inventory balance equations. Constraints (5) and (6) ensure that the correct pallet class pc is chosen and the corresponding transportation costs $TP_{dc,pc}$ are considered in the objective function. Equations (3) and (4) guarantee that only one pattern m of the feasible patterns $M_{f_{i,j}}$ is chosen and that the lot size $Q_{i,t}$ is positive if and only if this pattern provides a transport in t. Assuming a five day working week and a planning horizon T of four weeks, a feasible pattern would be Mon-Wed-Thu (recurring 4 times), whereas, Mon-Tue-Wed would not be allowed, since the supply of a lean production site should be equally distributed over the days of the week. Equations (7)-(9) restrict the total lot sizes (sum over all products) to be the same for all shipments.

This model is a basic formulation which was developed for the purpose of showing the fundamental functionality of the concept. In any case, for real-world use the model should and can be easily adapted to particular conditions and requirements on site (e.g. inventory or production capacity, minimal required frequency) and to the concrete business model of the network provider (e.g. tarif structure, level of allowed variation of lot sizes).

The solution of this model includes the optimal pattern and the minimal cost. We could show experimentally that costs are very insensitive to choosing another pattern with the same frequency. The cost of the optimal frequency is then evaluated by the average of the feasible patterns in $M_{f_{i,j}}$. Furthermore, a demand forecast

is, usually, afflicted with uncertainty. To generate a solution robust to changes of daily demand, a scenario approach was chosen. Thus, a fast solution method is crucial. Fortunately, the limited number of feasible patterns imposes a special structure that can be used to parameterise the branch-and-bound algorithm. First experiments showed, that instances with up to eight products and four weeks are solved with CPLEX 12.1 in less than 4 seconds (PC 2,50 GHz and 4 GB RAM).

Once all network clients have fixed their frequency bookings, the bookings are communicated to the network operator. Together with given fixed assignments of clients to transhipment points and transportation times, all inputs for the second stage problem are known. In this step the services are scheduled by mapping weekly patterns to frequencies in a way that the maximal number of pallets per day and region (for pick-ups and deliveries separately) are minimised.

Variables:	s_n^1, s_n^2	auxiliary variable for linearising the max operator of the objective function (with $n \in N$ the set of transhipment points, i.e. the set of pick-up and delivery regions)
	$x_{i,j,m}$	1 if pattern m is chosen for the transport from client i to j, 0 else (with $i, j \in C$)
	$y_{i,j,t}$	1 if there is a delivery on week day t from i to j, 0 else
	$q_{i,j}$	lot quantity from client i to j
Parameters:	$PAT_{m,t}$	week pattern m in time step t, i.e. $\{1,0,1,0,1\}$ for Mon-Wed-Fri (with $m \in M_{f_{i,j}}$ the set of reasonable patterns for the frequency $f_{i,j}$ and $t \in 1,\ldots,5$)
	$lz_{i,j}$	transport time from client i to j
	wd	number of working days
	R_n	set of clients in region n (with $n \in N$)

$$\min \sum_{n=1}^{N} (s_n^1 + s_n^2) \tag{12}$$

$$s.t. \quad \sum_{i \in R_n} \sum_{j} (x_{i,j,m} \cdot PAT_{m,t} \cdot q_{i,j}) \leqq s_n^1 \qquad \forall \quad t,n \tag{13}$$

$$\sum_{j \in R_n} \sum_{i} (y_{i,j,t} \cdot q_{i,j}) \leqq s_n^2 \qquad \forall \quad t,n \tag{14}$$

$$\sum_{m} x_{i,j,m} \cdot PAT_{m,t} = y_{i,j,[(t+lz_{i,j}) \bmod wd]} \qquad \forall \quad i,j,t \tag{15}$$

$$\sum_{m \in M_{f_{i,j}}} x_{i,j,m} = 1 \qquad \forall \quad i,j \tag{16}$$

$$x, y \in \{0,1\} \quad s^1, s^2 \geqq 0 \tag{17}$$

The auxiliary variables s_n^1 and s_n^2 along with constraints (13) and (14) form the linearisation of the maximum operators in the objective function. Feasible patterns $PAT_{m,t}$ are only considered by the pick-up volumes per day and region in (13), since the pick-up pattern of a relation (i,j) restricts the corresponding delivery pattern through constraints (15). With a five day working week and one day transportation time, the corresponding delivery pattern to Mon-Wed-Fri would be Mon-Tue-Thu. Constraints (16) ensure that every frequency booking is assigned to just one pattern of $M_{f_{i,j}}$.

For a state of the art MIP solver it takes around 8 minutes to solve the real-world instance presented in the following section to optimality (PC 2,50 GHz and 4 GB RAM, Gurobi 2.0.0). However, finding a feasible solution is easy: Randomly choosing a pattern for every booking, i.e. to fix one $x_{i,j,m}$ to 1 with $m \in M_{f_{i,j}}$ for every relation, produces a feasible solution. The idea for a simple heuristic is to use the information of the relaxed solution of a LP solver to perform this fixing in a more sensible way. This heuristic generates a solution in around three seconds for the same instance with an objective value 2% above the optimal one. We implemented this method for fast screenings of scenarios.

3 A Real World Application

We demonstrate the benefits of balanced transports on the network level with a scenario based on real-world[2] data obtained from a project partner – a major automotive manufacturer located in northern Germany. We compared the results of our optimisation to results, where feasible patterns were mapped randomly to frequency bookings.

The first column of Table 1 shows that the random module produces differences in daily pick-up volumes of up to 568 palletts and up to 40% relative to the mean value[3]. In case of the *LogoTakt* heuristic, the maximal absolute range is 29 pallets while the relative values – with the exception of Berlin – are less than 5%. The results for the corresponding delivery volumes per region are even better balanced.

Table 1 Differences in daily pick-up volumes and tours per region

Regions	number of pallets				number of tours				
	random		*LogoTakt*		random		*LogoTakt*		
	range	range/ mean	range	range/ mean	range	range/ mean	range	range/ mean	saved tours
Leipzig	408	19%	28	1%	16	19%	4	5%	1
Hagen	568	19%	11	0%	21	19%	3	3%	1
Hamburg	140	32%	21	5%	5	29%	2	12%	5
Hannover	145	5%	1	0%	6	6%	2	2%	3
Köln	457	30%	29	2%	17	29%	3	5%	9
Kornwestheim	134	9%	1	0%	5	9%	2	3%	1
Mannheim	257	12%	1	0%	10	12%	4	5%	4
Berlin	68	40%	18	11%	3	42%	1	14%	3
Nürnberg	263	23%	23	2%	10	22%	3	7%	2
München	215	21%	13	1%	10	25%	2	5%	12

The second half of the table shows the same indicators for the number of the resulting pick-up tours per region: again the ranges are significantly larger in case of the random mapping, whereas the driving distance was almost the same.

[2] The data set contains more than 3'500 relations and the network has 10 rail ports.

[3] The presented result is representative for the random mapping of bookings.

However, the most decisive figure for saving direct costs is, how much maximal capacity can be saved by avoiding volume peaks during the course of the week: The outmost column of table 1 shows, that the maximal number of tours of the random solution includes at least one tour more than the *LogoTakt* case. In total 41 pick-up tours (that is 7%) per week during a planning horizon of four weeks are saved – for the delivery tours a similar result can be achieved. The individual cost savings experienced by a network provider depend on the fleet structure.

4 Conclusions

The results show, that the introduced concept and the corresponding planning methods fulfil the requirements of a lean long-haul logistics network: The network loads are evenly distributed over the days of the week and for the duration of the planning period all transports recur weekly across company/production site lines. Transports are regular, transparent and easy to control, while the booking of transport frequencies corresponds to production frequencies. The results become even more interesting in an European network with more than one company incorporated. But, especially, in cases of low transport volumes per region and in the case of returnable empties (served by pick-up and delivery tours), it might be interesting to integrate the process of mapping weekly patterns to frequencies and solving the VRP.

References

1. H. Andersson, A. Hoff, M. Christiansen, G. Hasle, and A. Løkketangen. Industrial aspects and literature survey: Combined inventory management and routing. *Computers & Operations Research*, 37: 1515–1536, 2010.
2. L. Buschkühl, F. Sahling, S. Helber, and H. Tempelmeier. Dynamic capacitated lot-sizing problems: a classification and review of solution approaches. In *OR Spectrum*, volume 32, pages 231–361. Springer, April 2010.
3. T. Crainic. Long-haul freight transportation. In W. H. Randolph, editor, *Handbook of transportation science*, pages 451–516. Springer, 2003.
4. P. Francis, K. Smilowitz, and M. Tzur. The period vehicle routing problem and its extensions. In *The vehicle routing problem: latest advances and new challenges*, pages 73–102. B. Golden and S. Raghavan and E. Wasil, 2008.
5. U. Selinger, G. Liedtke, and M. Hinding. LogoTakt: high frequency delivery for medium flows of goods. In *Conference proceedings of the international trade and freight transportation conference ITFTC*, 2008.
6. J.P. Womack, D.T. Jones, and D. Roos. *The machine that changed the world.* Free Press, 2007.

Design of Less-than-Truckload Networks: Models and Methods

Jens Wollenweber and Stefanie Schlutter

Abstract We consider the design of a two-level transportation network. This type of network is typically established in national European less-than-truckload businesses. Cargo shipment from source to sink customers can be divided into three segments. Goods are picked up at customer locations and shipped to a source depot. After that there is a long-haul transport to a sink depot. Finally cargo is distributed from the sink depot to its final destination. In order to formalize this problem, two models, one linear and one quadratic, are proposed and analyzed. The goal of optimizing this NP-hard problem is to minimize total costs by determining the number, position and capacity of depots. In addition, customers have to be allocated to depots. To solve the optimization problem, we propose an approach which utilizes a relaxation of the long-haul transports. The resulting problem is the well-known facility location problem. Using this approach, real-world problems can be solved very quickly in high quality, for both an uncapacitated and a capacitated version of the model.

1 Introduction

High price pressure, newly opening markets, external effects such as tolls, as well as revised driving times and working hour regulations lead to ever increasing pressure for European transport service providers to optimize their business. In this article we examine a transport network that ships goods in a two stage handling process from source to sink customers. The transported goods typically have a weight between 50 and 3 000 kg. They are picked up at the sources, consolidated in a depot and preferably transported to the sink depot in a full single load direct transport. Here cargo is deconsolidated and delivered to the respective destinations. In this article

Fraunhofer Center for Applied Research on Supply Chain Services SCS, Department Networks, Nordostpark 93, 90459 Nuremberg, Germany e-mail: jens.wollenweber@scs.fraunhofer.de, stefanie.schlutter@scs.fraunhofer.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_42,

the design of such a network is viewed on a strategic level by optimizing number, location and capacity of depots, allocation of customers to the depots as well as the transports between the depots (main runs).
An overview of different network designs can be found in O'Kelly [4], Crainic [2, 3] and Wieberneit [7]. The literature found here essentially deals with models and procedures in a tactical planning horizon for which depotl location are already fixed. Strategic problems are mainly addressed in the field of facility location, see Aikens [1], Revelle and Laporte [6], as well as Owen and Daskin [5]. To our best knowledge, there is currently no scientific literature which addresses strategic network design.

2 Modeling

In the following, we introduce a two stage facility location/network design problem. In the basic version, U2NDPSS, unlimited capacity of the depots is assumed, while in a second version, C2NDPSS, the capacity of every depot is limited.

2.1 Uncapacitated Two Stage Network Design Problem with Single Source Restrictions (U2NDPSS)

First we address the uncapacitated problem. Between any two customers i, j there is a shipping quantity of a_{ij}. This quantity is handled first in an inbound depot k and second in an outbound depot l. This means a feasible transport can be broken down into three parts consisting of pick-up (source customer to inbound depot), main run (inbound to outbound depot) and delivery (outbound depot to sink customer). The depots as well as the transport routes are not limited in capacity. Furthermore it is necessary that for each customer both inbound and outbound depot are identical. Main run quantities are a direct result of the customer-depot allocation. For the operation of a depot k, fixed cost f_k have to be included. Shipping costs are linear dependant on the transport quantity. They can be separated into the three stages pick-up from customer i to depot k c^V_{ik}, main run from depot k to depot l c^H_{kl} and final-delivery from depot l to customer j c^N_{jl}.

U2NDPSS1 - Path formulation In the path formulation the transport between two customers i, j via the two necessary depots k, l are defined with a continuous variable $x_{ijkl} \in \mathbb{R}^+$. $z_k \in \{0,1\}$ has value 1, if the depot k is opened and 0 otherwise. The variable $y_{ik} \in \{0,1\}$ shows, if customer i is assigned to depot k ($:= 1$) or not ($:= 0$). With this, we get the following model:

$$\min \sum_k f_k z_k + \sum_{i,j,k,l} (c^V_{ik} + c^H_{kl} + c^N_{jl}) x_{ijkl}$$

$$\sum_{k,l} x_{ijkl} = a_{ij} \qquad \forall i,j \tag{1}$$

$$x_{ijkl} \leq a_{ij} y_{ik} \qquad \forall i,j,k,l \tag{2}$$

$$x_{ijkl} \leq a_{ij} y_{jl} \qquad \forall i,j,k,l \tag{3}$$

$$\sum_{k} y_{ik} = 1 \qquad \forall i \tag{4}$$

$$y_{ik} \leq z_k \qquad \forall i,k \tag{5}$$

Constraint (1) guarantees the flow conservation from customer i to customer j. With the constraints (2) and (3) the assignment of a customer i to a depot k is ensured for both pick-up and delivery. The allocation of customer i to exactly one depot k is realized with (4). If a customer i is connected to the depot k, then (5) ensures that the depot is open. The objective function minimizes fixed facility and transportation costs. The number of variables, in particular x_{ijkl}, strongly increases with the number of customers and depots.

U2NDPSS2 - Quadratic formulation In a second, quadratic formulation only the assignment of customer i to depot k is explicitly defined by variable y_{ik}. Due to the single source property, the main run quantities automatically result from the product of the assignment variables. For the shipping quantity between depots k and l, we have the additional variable $v_{kl} \in \mathbb{R}^+$. So we receive the model:

$$\min \sum_{k} f_k z_k + \sum_{i,j,k} a_{ij} c^V_{ik} y_{ik} + \sum_{i,j,l} a_{ij} c^N_{jl} y_{jl} + \sum_{k,l} c^H_{kl} v_{kl}$$

$$v_{kl} = \sum_{i,j} a_{ij} y_{ik} y_{jl} \qquad \forall k,l \tag{6}$$

with respect to restrictions (4) und (5). The new constraint (6) ensures that the total quantity from customer i to customer j is transported correctly. The model is more compact than the path formulation but the linearity is lost.

2.2 Capacitated Two Stage Network Design Problem with Single Source Constraints (C2NDPSS)

Due to several real-world constraints (i.e. number of gates, handling times), a model is introduced in the following that assumes a capacity limitation of the depots. With respect to the strategic model characteristics we assume a quantity related restriction. Therefore time aspects such as arrival times are not explicitly considered.
In chronological sequence the depots adopt different functions for either handling pick-ups or deliveries. For simplification we propose a capacity s_k which holds for both depot operating modes.
The path formulation of the capacitated model is carried out, by adding the follow-

ing restrictions to U2NDPSS1:

$$\sum_{i,j,l} x_{ijkl} \leq s_k z_k \qquad \forall k \tag{7}$$

$$\sum_{i,j,k} x_{ijkl} \leq s_l z_l \qquad \forall l \tag{8}$$

Inequality (7) limits the pick-up quantity assigned to depot k. For the outbound depot l the delivery quantity is limited in the same way by restriction (8).
The quadratic formulation of the capacitated model, C2NDPSS2, is modeled by replacing restriction (5) with the following restrictions:

$$\sum_{i,j} a_{ij} y_{ik} \leq s_k z_k \qquad \forall k \tag{9}$$

$$\sum_{i,j} a_{ij} y_{jl} \leq s_l z_l \qquad \forall l \tag{10}$$

Here (9) restricts the inbound quantity of depot k and (10) limits the outbound quantity of depot l. Like U2NDPSS, C2NDPSS is a NP-hard problem as well.

3 Relaxation and Approximation Based Solution Approach for U2NDPSS and C2NDPSS

Though the linear versions of both problems can be solved with standard MIP solvers, it cannot be expected that large instances can be solved efficiently. Therefore we propose a heuristic solution procedure and compare the performance with the results of a standard solver in the following section.

We propose a relaxation of U2NDPSS2 by setting all main run costs to zero. In the model we drop constraint (6) and set the last term of the objective function to zero. The objective function of this relaxation, U2NDPSS2-Rel-H, can be formulated as follows:

$$\min \sum_{k} f_k z_k + \sum_{i,j,k} \left(a_{ij} c^V_{ik} + a_{ji} c^N_{ik} \right) y_{ik}$$

under consideration of (4) and (5).
The resulting model is an uncapacitated facility location problem (UFLP) and can be solved in sufficient quality by standard MIP solvers. The optimal solution of U2NDPSS2-Rel-H is therefore a lower bound for the original network design problem.
Due to the location choice and the single assignment of customers to depots, the main run quantities are direct result of the customer-depots assignments. We receive an upper bound of the original problem by calculating the resulting main run costs and adding them to the lower bound.

In cases of high main run costs, the solution received by adding main run costs to the lower bound will not provide very good upper bounds. In order to improve the upper bound, we propose an iterative solution approach in which an approximation of the main run costs is added in the objective function of U2NDPSS2-Rel-H. These virtual main run costs are updated in every iteration. By that we generate a sequence of improving solutions, though it is not guaranteed that the solutions created by this approach converge to the optimum solution.

U2NDPSS3 - simplified model We propose a simplified version of U2NDPSS2 by replacing parameter c^H_{kl} with $\tilde{c}^H_k$. This new parameter approximates the long-haul costs, but in contrast to the original model it depends on the sink depot only. This modification leads to the following model:

$$\min \sum_k f_k z_k + \sum_{i,j,k} \left(a_{ij} c^V_{ik} + a_{ji} c^N_{ik} + a_{ij} \tilde{c}^H_k \right) y_{ik}$$

with respect to (4) and (5). In the capacitated version, constraints (9) and (10) have to be considered instead of (5).

In an iterative approach we solve U2NDPSS3 with changing main run costs: In the first iteration we compute the average main run costs of all depots by adding up all main run cost coefficients and dividing them by the number of depots:

$$\tilde{c}^H_k = \frac{\sum_l c^H_{kl}}{\sum_k 1}$$

After solving the U2NDPSS3 with the approximated costs, the real costs of the solution (upper bound) can be calculated easily since they are fully determined by the choice of depots and customer assignments. In the next iterations we update the approximated main run costs by the real average costs of the current solution:

$$\tilde{c}^H_k = \frac{\sum_{i,j,l} a_{ij} y_{ik} y_{jl} c^H_{kl}}{\sum_{i,j,l} a_{ij} y_{ik} y_{jl}}$$

Note that only depots are affected which are chosen in the current solution. Then U2NDPSS3 is solved again by using the new approximated costs. The algorithm is stopped if the gap between the best solution and the lower bound is below a certain limit or until we reach a maximum number of iterations. Though we cannot update the lower bounds, it is very likely that we can improve the upper bounds by this approach.

4 Computational Tests

The performance of the heuristic approach was tested with data set of 75 customers and 45 depots and compared to the results of the MIP solver IBM ILOG CPLEX. From this data set, 32 instances with different pickup/delivery, main run and fixed depot costs are derived. Additionally, we define three different capacity limits for the C2NDPSS. We have limited computation time to 3 hours.
For the U2NDPSS both the heuristic approach and CPLEX were able to find solutions for every of the 32 instances. The average time for the heuristic approach is 10 seconds, while CPLEX takes 74.5 minutes. The heuristic approach computes solutions which are on average 17.7% better than the MIP results. For 14 instances CPLEX provided a slightly better solution (maximum 0.78%), while for 18 instances the heuristic approach provides a better solution (up to 57.46%).
For the C2NDPSS CPLEX was not able to calculate a feasible solution for any instance within the given computation time. The heuristic approach again provides very quickly (avg. 8.5 seconds) solutions.

5 Conclusion and Further Research

In this paper we introduce models for strategic network design. We propose an iterative solution approach based on the relaxation of the main run. In first computational tests it could be shown that this approach outperforms a standard MIP solver by far. By solving the proposed models, we can get a first idea how to locate (and size) depots. In order to solve real-world problems, it is necessary to consider additional constraints, for instance synchronisation constraints in the depots. This will be subject of further research.

References

1. C. H. Aikens. Facility location models for distribution planning. *European Journal of Operational Research*, 22: 263–279, 1985.
2. Teodor Gabriel Crainic. Service network design for freight transportation. *European Journal of Operational Research*, 122: 272–288, 2000.
3. Teodor Gabriel Crainic and Gilbert Laporte. Planning models for freight transportation. *European Journal of Operational Research*, 97: 409–438, 1997.
4. Morton E. O'Kelly and Harvey J. Miller. The hub network design problem – a review and synthesis. *Journal of Transport Geography*, 2(1): 31–40, 1994.
5. S. H. Owen and M. S. Daskin. Strategic facility location: A review. *European Journal of Operational Research*, 111: 423–447, 1998.
6. C. S. Revelle and G. Laporte. The plant location problem: New models and research prospects. *Operations Research*, 44: 864–874, 1996.
7. Nicole Wieberneit. Service network design for freight transportation: a review. *OR Spectrum*, 30: 77–112, 2008.

A Comparison of Recovery Strategies for Crew and Aircraft Schedules

Lucian Ionescu, Natalia Kliewer, and Torben Schramme

Abstract On the day of operations delays frequently forces the operations control of an airline to recover resource schedules by mostly expensive recovery actions. Regarding this difficulty robust scheduling deals with the construction of schedules which are less affected by disruptions or provide possibilities for low-cost recovery actions. To measure the degree of robustness of schedules, a simulation framework with enabled recovery actions is needed. In this paper we present a comparison of re-optimization and rule-based recovery approaches.

1 Introduction

On the day of operations the crew and aircraft schedules may become infeasible due to disruptions, e.g. the unavailability of aircraft or crew, bad weather conditions or missing passengers. These disruptions lead to infeasible schedules which have to be recovered at the operations control. The process of finding new schedules is called recovery or disruption management. To recover a disrupted schedule airlines use actions like aircraft and crew swapping, rerouting passengers, using reserve or standby resources, cancelling or delaying flights, until all resources for operation are available. During the recovery process several restrictions has to be respected, e.g. preservation of legality of crew pairings and also adherence of maintenance intervals for aircraft. [7] present an outline of fundamental recovery techniques in airline operations, whereas [2] give a comprehensive overview of state-of-the-art concepts and models for recovery of different airline resources, focusing on the computational aspect of recovery. Delays, cancellations and other recovery actions lead to

Lucian Ionescu · Natalia Kliewer
Information Systems, Freie Universität Berlin, Garystraße 21, 14195 Berlin, Germany, e-mail: {Lucian.Ionescu, Natalia.Kliewer}@fu-berlin.de

Torben Schramme
Decision Support & Operations Research Lab, Universität Paderborn, Warburger Straße 100, 33098 Paderborn, Germany, e-mail: schramme@mail.upb.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_43,

additional cost, which can be reduced if the resource schedules are either likely to remain feasible under disruptions or can be easily recovered when disruptions occur. This general approach is called robust scheduling. The goal of robust scheduling is to minimize both planned and reactionary costs caused by disruptions. Thus, robust scheduling can lead to higher planned costs compared with cost-efficient scheduling, however, this will be compensated by savings in reactionary costs.

As considering robustness we distinguish between two aspects, namely stability and flexibility Stability determines the degree of ability of a schedule to remain feasible despite unpredictable disruptions. For stability in airline scheduling see for example [9], [6] or [5]. In contrast, flexibility deals with the degree of ability to react to resource allocation conflicts by adapting the originally planned schedule to the operating environment. Examples for flexible scheduling are [8] and [1]. Under consideration of real world requirements, flexibility means that schedules can be recovered by *simple* and preferably local actions, so that the schedule modifications do not spread over large parts of the schedule. Thus, it is desirable to construct schedules that tend to remain feasible or can be recovered by selective local recovery actions.

In this paper we measure the performance of different basic approaches for the crew and aircraft recovery problem. We compare a online re-optimization approach with a rule-based strategy. Our objectives are the punctuality of flight arrivals and the run-time efficiency. For the punctuality we compute a lower bound by offline re-optimization, an upper bound can be obtained by using delay propagation only.

2 Robustness Measures

A practicable way to measure robustness and thus operational costs of resource schedules is the on-time performance (OTP) of operated flights in a simulation environment. OTP is the fraction of flights arriving on-time at their destination airport. The inclusion of additional delay tolerance thresholds, e.g. 0, 5 and 15 minutes, is also possible. Hence, a flight is on-time regarding a 15 minutes threshold if its arrival delay does not exceed 15 minutes. The OTP can be increased by stable planning and by recovery on the day of operations. Other measures for robustness are the number of rule violations for crew duties or for aircraft maintenance intervals or the overall amount of all propagated delays within the flight network. We take into account both stability and flexibility aspects by combining the delay propagation with further recovery actions. Thus, delays are propagated until they are resolved by schedule modifications or compensated by buffer times. Delay propagations between crew duties are allowed, although this should be prevented in order to avoid a delayed beginning of the next working period.

Due to the fact that recovery actions are a common instrument in airline practice, the consideration of recovery actions when evaluating robust schedules is more preferable than using delay propagation only, which would lead to underestimation of schedule robustness. Using only delay propagation when simulating schedule operations delivers an upper bound for the robustness of schedules.

Due to the online nature of the recovery problem, finding good recovery strategies is a big challenge because new, recovered solutions must be found very quickly. There exist a lot of sophisticated re-optimization approaches for the online resource re-scheduling problem in the literature, see [2]. Re-optimization strategies allow to modify the whole upcoming schedule starting from the time information about a future delay becomes available. The alternative to the re-optimization are the rule-based recovery strategies, which mostly resolve arising conflicts following ex ante defined rules or heuristics. Furthermore rule-based strategies operate mostly locally at the point of the network, where a disruption occurs. We assume that rule-based strategies are in general less time-consuming.

Regarding the evaluation of robust schedules the rule-based strategies offer another positive aspect. As we want to create schedules that are able to be recovered easily, a re-optimization including the possibility to reschedule all upcoming tasks may falsify the evaluation, because in the current practice schedules would never be modified to this extent. Thus, an OTP resulting from re-optimization recovery is an insufficient indicator for the schedule robustness; in worst case it presents only an indicator for more or less good recovery. Local recovery strategies avoid this, as it is our intention to create schedules that can be recovered by a manageable number of local actions. In this regard another significant aspect is the running time. An evaluation carried out by simulation claims for fast recovery due to the running time of hundreds of simulation runs with hundreds of primary delays.

In this paper we want to estimate the gap between re-optimization and rule-based recovery by two ways. First, we compare the online re-optimization approach with a theoretical offline re-optimization approach (ex-post) where all disruptions of underlying scenarios are known in advance. In contrast, the online approach is only able to use presently available delay information. Consequently, the offline approach is not a practicable recovery technique but gives us a theoretical lower bound for the recovery problem. The result of this evaluation reveals the *online-offline* gap of the recovery problem.

The second issue of this paper is the comparison between online re-optimization and online rule-based recovery. On the one hand, we want to find out if there is a loss of OTP when not using re-optimization. On the other hand, we want to analyze, if the gap between both recovery strategies is constant. A constant gap enables rule-based strategies for schedule evaluation because more robust schedules perform better than less robust schedules. All strategies solve the integrated recovery problem for both crews and aircraft. Besides the comparison of the OTP, we also consider the running times of the strategies.

3 Schedule Recovery Strategies

For the offline analysis as well as for the online re-opt strategy, we use the iterative recovery method with retiming by [3]. This approach uses a column generation framework to solve the aircraft routing and the crew pairing problem iteratively. During the search for a new feasible solution for the current scenario ω, it is possible

to retime, i.e. to delay, upcoming flights. In doing so, this re-optimization includes swapping, cancellations (uncovered tasks) and delay propagation (by retiming) implicitly.

In contrast, the rule-based strategy works as follows: if a delay is below a configurable delay tolerance threshold t_d, e.g. 15 minutes, it is directly propagated without looking for possible recovery actions. If it exceeds t_d, we check if the delayed flight has to be cancelled, i.e. if a flight is delayed by more than threshold t_{canx}. For us, cancellations are only resource based decisions, incorporating only crew and aircraft routes, passenger reroutings are not considered. If a flight is cancelled, the whole following cycle until the crew and aircraft shows up at the current airport has to be cancelled, too. Cancelled flights are classified as not on-time with regard to all OTP steps. If no cancellations will take place (and that is in most cases), we search for a swap opportunity for both crew and aircraft. A possible swap preferably swaps crew and aircraft to the same new successor task. If there is no identical swap possible, crew and aircraft may be swapped to different new assignments. [8] formulate the demand that the resources are swapped back to their original routes later on. We use both possibilities – crew and aircraft can be swapped back later or finish their new itinerary if it is feasible regarding crew regulations and maintenance intervals. If there exists no swap possibility, the delay is propagated to successor tasks of crew and aircraft.

4 Computational Results

For our analysis of both online and offline recovery strategies we use given flight and aircraft schedules of an European airline in an event-based simulation framework. The schedules include up to 200 flights each and last from one to three days. Based on the aircraft rotations, we constructed crew schedules with different stability degrees by a local indicator based approach, presented by [9]. The optimization framework is described in [4]. Different delay scenarios $\omega \in \Omega$ with exponential distributed delays are randomly generated. All computational experiments were carried out on an Intel Core i7 (2.86GHz) processor with 4096 MB RAM. As we compare the rule-based recovery with propagation-only, we can differ between the effect of stable planning only and the effect of incorporated recovery during simulation. The higher the penalty for non-stability, the more stable the schedules get. Even though we increase the stability of schedules, the additional OTP achieved by recovery actions is almost constantly about 6% to 7% (see figure 1). Thus, increasing stability does not imply an eminent decrease of flexibility. Further on, and that is an interesting aspect, enabling the rule-based approach for evaluation of schedules - the rule-based approach maintains a constant distance to the online and offline approach, respectively. Thus on the one hand the local rule-based approach covers most of all possible recovery actions, on the other hand the loss of possible recovery actions by re-optimization remains constantly. This enables us to use the run-time saving rule-based approach for evaluation of schedules, because the results are not getting falsified although we underestimate the on-time-performance by a nearly constant

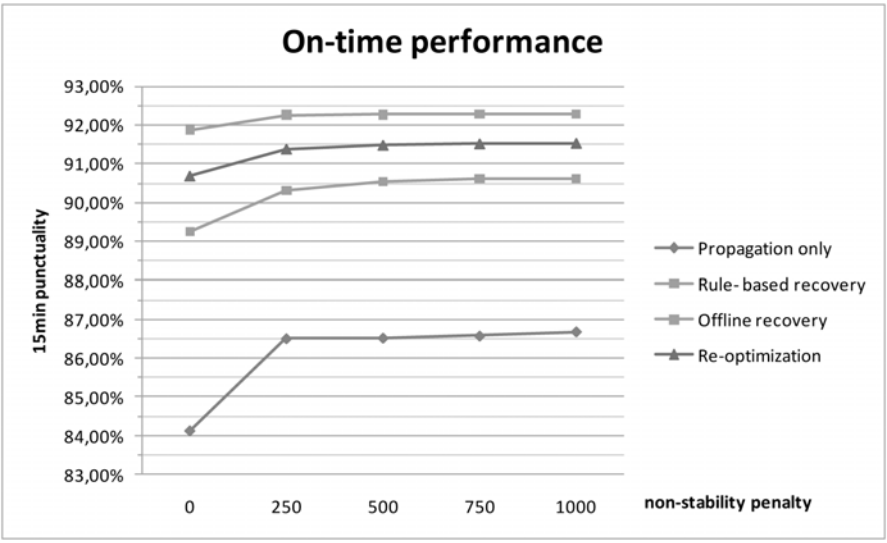


Fig. 1 OTP of crew and aircraft schedules with different recovery settings

factor. Besides it is interesting to see that the offline approach - giving us a theoretical upper bound for the OTP - does not increase the OTP significantly, when we use more stable schedules. The online recovery approaches do not show this effect, the OTP increases slightly for all settings but the offline approach. This observation shows that the overall online-offline gap of the OTP decreases by increasing stability.

Table 1 Run-time of different recovery strategies for a selection of flight schedules [in seconds]

strategy	199	196	132	110	104	116	83	96
re-optimization	58682	69806	9701	18459	8515	4969	1523	9584
offline recovery	6348	6572	6652	2105	737	1456	306	1153
rule-based	3	3	3	2	2	2	2	2
propagation only	3	3	2	2	2	2	2	2

Discussing the run-time of all strategies it is obvious that the rule-based strategies do perform considerably better (see table 1). They only need a small fraction of the run-time of the re-optimization based approaches. In connection with the constant OTP gap, the rule-based approach is favourable for evaluation of schedules. The offline re-optimization is faster than the online re-optimization due to the fact, that the offline version only has to carry out one optimization per delay scenario, because all delays has to be known. Online re-optimization is carried out after each upcoming delay.

5 Summary and Outlook

The main contribution of this work is to show that a local rule-based recovery approach enables us to evaluate the robustness of airline resource schedules without falsifying the results, although we underestimate the on-time-performance by a constant factor. Future research will have to deal with consideration of increasing both stability and flexibility in crew and aircraft scheduling. In particular, it will be interesting to see in what extent stability and flexibility affect each other. So far, our results give a slight remark that a higher degree of stability does not urgently decrease the possibilities to recover schedules from occurring disruptions. This can be explained by the fact, that longer buffer times may also increase the number of possible swaps of resources to a certain extent. However, this assumption still needs to be definitely verified incorporating both stability and flexibility in scheduling.

References

1. Y. Ageeva. Approaches to incorporating robustness into airline scheduling. Master's thesis, Massachusetts Institute of Technology, August 2000.
2. J. Clausen, A. Larsen, J. Larsen, and N. Rezanova. Disruption management in the airline industry – concepts, models and methods. *Computers & Operations Research*, 37: 809–821, 2010.
3. V. Dück, T. Schramme, N. Kliewer, and L. Suhl. An iterative approach for recovery of airline crew and aircraft schedules. Technical report, Decision Support & Operations Research Lab, University of Paderborn, Warburger Straße 100, 33098 Paderborn, Germany, 2009.
4. V. Dück, F. Wesselmann, and L. Suhl. A branch-and-price-and-cut framework for crew pairing optimization. Technical report, Decision Support & Operations Research Lab, University of Paderborn, Warburger Straße 100, 33098 Paderborn, Germany, 2009.
5. J. R. Birge and J. W. Yen. A stochastic programming approach to the airline crew scheduling problem. *Transportation Science*, 40(1): 3–14, 2006.
6. S. Lan, J.-P. Clarke, and C. Barnhart. Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. *Transportation Science*, 40(1): 15–28, February 2006.
7. J. Larsen, A. Ross, S. Tiourine, N. Kohl, and A. Larsen. Airline disruption management – perspectives, experiences and outlook. *Journal of Air Transportation Management*, 13: 149–162, 2007.
8. S. Shebalov and D. Klabjan. Robust airline crew pairing: Move-up crews. *Transportation Science*, 40(3): 300–312, August 2006.
9. Oliver Weide, David Ryan, and Matthias Ehrgott. An iterative approach to robust and integrated aircraft routing and crew scheduling. *Computers & Operations Research*, 37(5): 833–844, May 2009.

A Case Study for a Location-Routing Problem

Sebastian Sterzik, Xin Wang, and Herbert Kopfer

1 Problem Analysis

In contrast to the classical location-allocation models, the location-routing problem (LRP) integrates vehicle routing and scheduling aspects in the location planning process. A survey and classification scheme of location-routing is e.g. proposed by [3] and [2]. This paper presents a LRP of a company which distributes finished cars. This LRP is decomposed into two steps. In the first step a set of candidate depots is determined by minimizing the sum of the radial distances between depots and customers. Based on the locations found in the first step the vehicle routing problem is used to select the depot with the most attractive location in the second step. This paper is organized as follows. In the following the problem is discussed followed by a solution approach in Section 2. The computational results are presented in Section 3. Finally, the main conclusions of the study are drawn in Section 4. The core business of the considered company is the transportation of finished vehicles from automotive factories to car dealers in different regions of Germany. The company's distribution network is structured as a Hub&Spoke network with each hub serving the customers within a defined region. This paper focuses on the interregional transportation of cars which will be necessary if a car dealer from one region demands cars of the type which are produces by an automotive factory in the second region. In the current situation interregional transports have to pass both terminals (see Fig. 1(1)). Hence, this interregional transportation proceeding goes along with unnecessary expense, especially if there exists an unbalanced demand between the observed regions. Due to this imbalance the capacity of trucks returning from the one region to the other region is lower at large and leads to a low degree of capacity utilization. Furthermore this proceeding goes along with negative effects for the truck drivers since they are not able to return to their initial point on the same day.

Chair of Logistics, Department of Economics and Business Studies, University of Bremen, Wilhelm-Herbst-Str. 5, 2359 Bremen,Germany, e-mail: {sterzik, xin.wang, kopfer}@uni-bremen.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_44,

Thus, a great drawback of the current distribution system constitutes the required tours between the terminals.

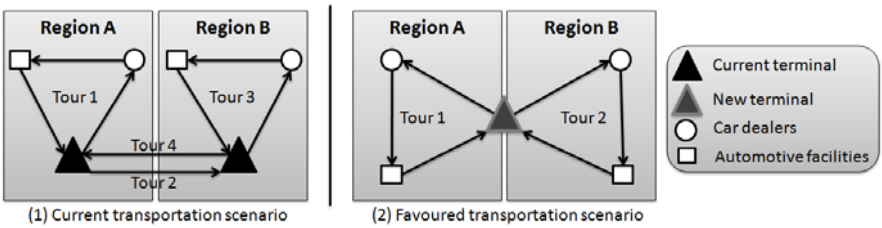


Fig. 1 The current and favoured interregional transportation scenarios

An additional terminal linking both regions can be used for interregional transportation in order to establish a more efficient scenario (see Fig. 1(2)). Thereby tours between the current terminals will be needless. Cars from automotive facilities in one region will be carried to car dealers in the second region in only two tours via the new terminal. The current terminals then solely are used for the transportation within their region. A demand imbalance between the regions is of little significance since trucks do not have to return to their home depot as in the current interregional transportation scenario. The optimization problem considered in this paper is the LRP for the additional interregional terminal.

2 A Sequential Location-Routing Method

The solution process for the LRP is divided into the following two steps:

1. In the first step potential depot locations are determined by means of the Fermat-Weber problem on the one hand and by considering the traffic infrastructure on the other hand. The Fermat-Weber problem is based on a Cartesian coordinate-system and can be solved by Miehle's algorithm ([1]). The attained solution minimizes the Euclidean distances to m fixed customers with distances weighted by the customers' supply and demand amount. Since this does not consider the traffic infrastructure the highways around the attained location should be taken into account, too. Thus, all cities nearby and those close to the highways which connect the considered regions constitute possible candidate locations for the new additional terminal. A diversified cross selection of $q \geq 3$ cities should give a suitable initial network for the second step.

2. In the second step the possible depot locations are rated on the basis of vehicle routing and scheduling results. Hence, implementing a realistic virtual distri-

bution scenario for each location and choosing the candidate with the lowest expected operational transportation costs will lead to the new terminal location.

The company's vehicle routing problem can be described by means of the vehicle routing problem with backhauls (VRPB) and the split-delivery vehicle routing problem (SDVRP). On each route, all deliveries must be completed before any pickups are allowed. Each customer can be visited more than once since the demand and supply amount of a customer may exceed the vehicle capacity. Since the route length is limited by means of the driving time restrictions the distances between the vertices are described by driving minutes. From the depot 0 all linehaul customers $L = \{1,\ldots,n\}$ with a certain car demand d and all backhaul customers $B = \{n+1,\ldots,n+m\}$ with a certain car supply f should be served. Thus, the problem can be defined over an undirected graph $G = (V,E)$ with vertex set $V = \{0\} \cup L \cup B$ and edge set $E = \{(i,j), i,j \in V\}$. An unlimited number of vehicles each with a capacity Q is available. Each vehicle must start and end its route at the depot. x_{ijk} denotes a Boolean variable equal to 1 if vehicle k travels directly from i to j and equal to 0 otherwise. Moreover y_{ik} defines the quantity of the demand/supply of customer i served by vehicle k. The objective is to minimize the total time distance t_{ij} to all customers in both regions.

$$\min z = \sum_{i=0}^{n+m} \sum_{j=0}^{n+m} \sum_{k=1}^{K} t_{ij} x_{ijk} \tag{1}$$

$$\sum_{j=0}^{n+m} x_{0jk} = \sum_{i=0}^{n+m} x_{i0k} = 1, \qquad k = 1,...,K \tag{2}$$

$$\sum_{i=1}^{n} y_{ik} \leq Q, \qquad k = 1,...,K \tag{3}$$

$$\sum_{i=n+1}^{n+m} y_{ik} \leq Q, \qquad k = 1,...,K \tag{4}$$

$$\sum_{k=1}^{K} y_{ik} = d_i, \qquad i = 1,...,n \tag{5}$$

$$\sum_{k=1}^{K} y_{ik} = f_i, \qquad i = n+1,...,n+m \tag{6}$$

$$y_{ik} \leq d_i \sum_{j=0}^{n+m} x_{ijk}, \qquad i = 1,...,n; k = 1,...,K \tag{7}$$

$$y_{ik} \leq f_i \sum_{j=0}^{n+m} x_{ijk}, \qquad i = n+1,...,n+m; k = 1,...,K \tag{8}$$

$$\sum_{i=n+1}^{n+m} \sum_{j=1}^{n} \sum_{k=1}^{K} x_{ijk} = 0 \tag{9}$$

$$\sum_{i=0}^{n+m}\sum_{j=0}^{n+m} t_{ij}x_{ijk} \leq T, \qquad k = 1,...,K \tag{10}$$

$$\sum_{i=0}^{n+m} x_{ilk} - \sum_{j=0}^{n+m} x_{ljk} = 0, \qquad l = 0,1,...,n+m; k = 1,...,K \tag{11}$$

$$u_{ik} - u_{jk} + 2 * Q \times x_{ijk} \leq 2 * Q - 1, \qquad i,j = 1,...,n+m; k = 1,...,K \tag{12}$$

$$1 \leq u_{ik} \leq 2 * Q \qquad i = 1,...,n+m; k = 1,...,K \tag{13}$$

$$x_{ijk} = \{0,1\}, \qquad i,j = 0,1,...,n+m; k = 1,...,K \tag{14}$$

$$y_{ik} \geq 0, \qquad i = 1,...,n+m \tag{15}$$

$$d_i \geq 0, \qquad i = 1,...,n \tag{16}$$

$$f_i \geq 0, \qquad i = n+1,...,n+m \tag{17}$$

Constraints (2) impose that every tour has to begin and end at the depot. Constraints (3)-(8) concern the allocation of the demand and supply amount of the customers among the vehicles. While (5) and (6) assure that the entire demand or supply of each vertex is satisfied, (3) and (4) ensure that the quantity delivered or picked up does not exceed the vehicle capacity. Constraints (7) and (8) impose that customer i can only be served by vehicle k if k passes through i. (9) guarantees that a truck does not drive to a car dealer after he visited an automotive facility. The abidance of the route length is ensured by restriction (10). Constraints (11)-(13) are the classical routing constraints; constraints (11) impose the continuity of a route while (12) and (13) guarantee that each vehicle performs a Hamiltonian cycle without exceeding the capacity limit of Q cars.

3 Computational Results

The regions Southwest Germany (SW) and Southeast Germany (SE) are known to have a significant demand imbalance. In the underlying request data set there were about 50% more cars which had to be moved from SE to SW than vice versa. Therefore, establishing a new terminal between these regions is of great interest for the company. In the first step of the sequential location-routing method we identified four possible candidate locations A-D (Fig. 2).

For the investigation of realistic distribution processes for each possible location we assume the amount of the supply and demand in the regions and construct daily routes to both regions. A truck cannot move more than eight finished vehicles at once. An additional important restriction for the company is defined by the legal driving time regulations. The EC regulations generally allow a truck driver to drive 540 minutes per day. Due to the operating experiences of the analyzed company this route length restriction is reduced further by a buffer of 40 minutes to consider

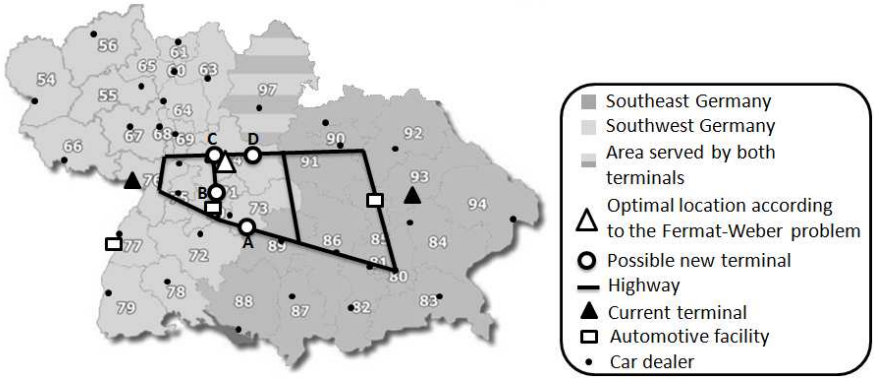


Fig. 2 Possible depot locations

possible time lags (e.g. traffic jams). The average demand and supply amount in SE and SW is shown in Table 1.

Table 1 Average amount of demand and supply in Southeast and Southwest Germany

Southeast Germany																
Postcode region	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	97
Demand	-	-	2	-	-	5	-	1	2	2	-	-	-	20	-	1
Supply	-	-	-	-	-	3	-	-	-	-	-	-	-	63	-	-

Southwest Germany																						
Postcode region	54	55	56	60	61	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	97
Demand	2	-	3	-	3	8	5	4	6	3	2	2	-	2	2	3	5	2	3	4	2	2	3
Supply	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-	7	22	-	-	-

Comparing the results of the distribution scenarios in SE and SW yields the rating of the locations. Location C provides the best results and is determined as the potential new terminal. At this point the found new solution with an interregional terminal at location C should be compared with the current situation. Therefore, the current scenario as seen in Fig. 1(1) is analyzed by means of the stated model above. The terminals in both regions serve as the depot for the distribution in SE and SW. Furthermore the connections between these depots have to be added to illustrate the interregional transportation proceeding. Table 2 shows the comparison between the current transportation net and the proposed net with the additional terminal C. Hereby the values are optimal for SE while the solution process for SW has been stopped when an optimality gap of 10% had been reached. For the current route planning the driving time within SE and SW is less than for the route planning with the new terminal since the current terminals are positioned more centrally in the regions. However, like stated in Section 1, the connections between the terminals have to be taken into account. Altogether the current solution leads to 15% additional total driving time compared to the route planning with location C as the new terminal.

Furthermore the new solution with an interregional terminal reduces the number of tours by about 28%.

Table 2 Comparison between the current and favoured distribution proceedings

Region	Current route planning (Driving time/ Number of tours)	Favoured route planning with terminal C (Driving time/ Number of tours)
Southeast Germany	2222/ 5[1]	6000/ 13[1]
Southwest Germany	3530/ 9[2]	3863/ 10[2]
Connection between the current terminals	2*2889/2*9	—
Total	11530/ 32	9863/ 23

[1] Optimally solved [2] Optimality Gap:10%

4 Conclusions

The presented LRP is solved by a method for combining location planning and vehicle routing and scheduling in a sequential way. The case study has shown that the proposed method can lead to significant savings in distribution costs. Providing reliable request data for solving the vehicle routing problem is essential to get a good solution. Nevertheless, reflecting a realistic distribution of goods within the location planning process can provide the opportunity to obtain solutions of higher quality than the classical location-allocation models.

Acknowledgements This research was supported by the German Research Foundation (DFG) as part of the Collaborative Research Centre 637 "Autonomous Cooperating Logistic Processes - A Paradigm Shift and its Limitations" (Subproject B7).

References

1. W. Miehle. Link-Length Minimization in Networks. *Operations Research*, 6(2): 232–243, March–April 1958.
2. H. Min. Combined location-routing problems: A synthesis and future research directions. *European Journal of Operational Research*, 108(1): 1–15, July 1998.
3. G. Nagy and S. Salhi. Location-routing: Issues, models and methods. *European Journal of Operational Research*, 177(2): 649–672, March 2007.

Planning and Control of Automated Material Handling Systems: The Merge Module

Sameh Haneyah, Johann Hurink, Marco Schutten, Henk Zijm, and Peter Schuur

Abstract We address the field of internal logistics, embodied in Automated Material Handling Systems (AMHSs), which are complex installations employed in sectors such as Baggage Handling, Physical Distribution, and Parcel & Postal. We work on designing an integral planning and real-time control architecture, and a set of generic algorithms for AMHSs. Planning and control of these systems need to be robust, and to yield close-to-optimal system performance. Currently, planning and control of AMHSs is highly customized and project specific. This has important drawbacks for at least two reasons. From a customer point of view, the environment and user requirements of systems may vary over time, yielding the need for adaptation of the planning and control procedures. From a systems' deliverer point of view, an overall planning and control architecture that optimally exploits synergies between the different market sectors, and at the same time is flexible with respect to changing business parameters and objectives is highly valuable. An integral planning and control architecture should clearly describe the hierarchical framework of decisions to be taken at various levels, as well as the required information for decisions at each level, e.g., from overall workload planning to local traffic control. In this research, we identify synergies among the different sectors, and exploit these synergies to decompose AMHSs into functional modules that represent generic building blocks. Thereafter, we develop generic algorithms that achieve (near) optimal performance of the modules. As an example, we present a functional module from the Parcel & Postal sector. In this module, we study merge configurations of conveyor systems, and develop a generic priority-based real-time scheduling algorithm.

Sameh Haneyah
University of Twente, 7500 AE Enschede P.O. Box 217, The Netherlands, e-mail: s.w.a.haneyah@utwente.nl

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 281
DOI 10.1007/978-3-642-20009-0_45,

1 Introduction

In our effort to identify synergies among the AMHSs used in different business sectors, we decompose AMHSs into functional modules that represent generic building blocks. Thereafter, we develop generic algorithms that achieve (near) optimal performance of the modules. Few authors address generic planning and control of AMHSs, Boyd et al. [2] review general control forms. Chen et al. [1] present a control framework for AMHSs using the holonic approach, however it does not solve detailed operational material handling problems. In this paper, we focus on the merge module, which appears in different types of AMHSs using conveyor belts (see Figure 1). Merge configurations consist basically of several infeed conveyors (or *infeeds*) that transport items to a *merge conveyor* on which the items are gathered. We develop a generic real-time scheduling algorithm that allocates space on the merge conveyor to items transported by the infeeds. In literature, studies related to merge configurations are more analogous to traffic systems than conveyor belts (see Shladover [7]). Studies relevant to conveyor systems are mostly simulation-based (El-Nashar and Nazzal [6]), where scheduling and control are not evident beyond simple routing rules. Many studies deal with AMHSs as a supporting resource within a manufacturing environment (see Bozer and Hsieh [3]).

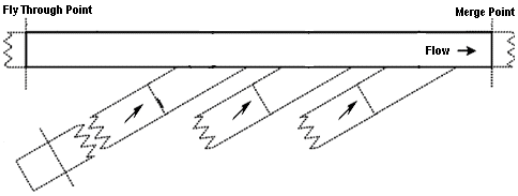


Fig. 1 Merge configuration.

The objective of the algorithm is two-fold: throughput maximization, and workload balancing among infeeds. Briefly stated, throughput refers to the output of the merge area measured in the average number of items passing per hour at the *merge point* (see Figure 1). This is equivalent to space utilization on the merge conveyor. Workload balancing refers to an even distribution of workload among all infeeds, this is reflected by synchronizing waiting times on the infeeds. As items on the infeeds look for merge spaces on the merge conveyor, space should be allocated in a way that results in an even distribution of waiting time elapsing before the items are physically merged. Current practices show that it is possible to optimize for one of these objectives, but at the cost of deteriorating the other objective. Therefore, it is important to maintain the right balance between the two conflicting objectives. In general, throughput is the main objective, and it is mostly unacceptable to optimize for workload balancing at the cost of low throughput. The main research question for this sub-problem is: How can the space on the merge conveyor be allocated to the parcels on the infeeds to achieve high throughput and a workload balance among

infeeds, and how to keep the right balance between these conflicting objectives? A relevant example from practice for this problem comes from the parcel & postal business sector. Figure 2 shows the generic layout of an express parcel sorting system. At the unload area, operators place parcels onto the infeeds. These infeeds transport parcels to the merge conveyor represented by the big loop in Figure 2. The merge occurs when the parcels transported on the infeeds reach the merge conveyor. After parcels are transferred to the merge conveyor, they circulate on the merge conveyor to reach the load area. There, parcels are automatically sorted to their destinations. Parcels are released into sorting *chutes* (see Figure 2). These chutes end at big containers where parcels with the same destination are collected in preparation for further transport. *Flying through* parcels are those parcels that flow back into the unload area, because their destination chutes were full or had some disruptions. McWilliams [4, 5] works on the parcel hub scheduling problem, but focuses on loading and unloading schedules, where the AMHS is dealt with as a black box .

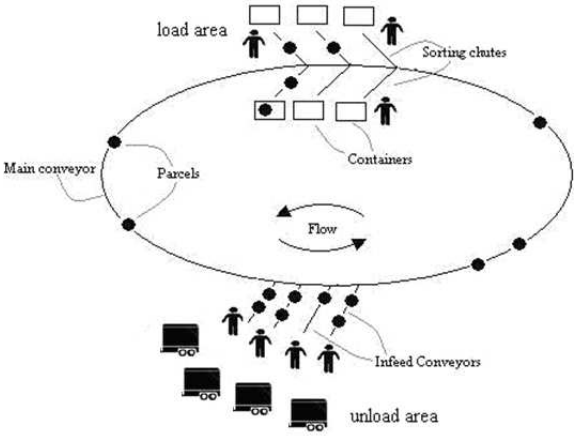


Fig. 2 Generic layout of a closed-loop conveyor sorting system.

We focus on this particular example. Section 2 illustrates the challenge of workload balancing that arises due to the early reservation phenomenon, and presents the general idea of the algorithm we develop to control the merge operation. Section 3 reports on some experimental results. Finally, Section 4 concludes this paper.

2 Early Reservation Phenomenon and Algorithmic Design

Figure 3 sketches a merge configuration with two infeeds of the same length; each infeed has a parcel. As soon as a parcel is loaded on an infeed, it is announced in the system and requests a merge space. The parcel from infeed 1 can arrive at point A, and therefore can reserve it. The parcel from infeed 2 can arrive at point

B. However, if all infeeds are busy with parcels, then already at an earlier decision moment, point B could have been allocated to some previous parcel from infeed 1. This phenomenon induces the dedication of most of the space as required by parcels from infeed 1, while forcing parcels from infeed 2 to wait for a space at a later point than requested. The main point is that parcels from infeed 1 can reserve spaces on the merge conveyor earlier than parcels from infeed 2, due to the restricted look ahead horizon. Therefore, as the system operates for long time, the total waiting time for parcels of infeed 2 accumulates. Moreover, when there is a larger system with more infeeds, this phenomenon propagates, and may result in high imbalance measures. The phenomenon occurs mainly when infeeds farther downstream are not long enough to see all incoming parcels (that compete for the same merge space) at the time of making allocation decisions. Then, parcels from those infeeds are forced to wait before being merged, more than parcels from upstream infeeds.

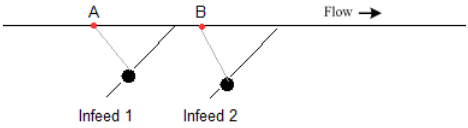


Fig. 3 Free space search.

We take a dynamic approach to the space allocation problem in merge configuration. We develop a generic real-time priority-based algorithm (PBA). The generic PBA contains three main processes:

- Process A: Merge space search process: This process is a simple search for an empty space on the merge conveyor. An announced parcel (on an infeed) may activate this process to search for an available merge space.
- Process B: The pending requests queue process: this process allocates empty space, appearing at the merge conveyor, to parcels waiting to be allocated. It uses priorities in allocating available space to waiting parcels. Priorities are calculated according to a formula that gives weights to space utilization and workload balancing. This process is activated by empty space appearing in the merge area.
- Process C: The reallocation process: this process aims at balancing the workload among infeeds by changing allocation decisions after more information becomes available in the system. This is an improvement procedure that is activated by (*priority*) parcels and not by empty space. Some parcels may overrule the space allocated to other parcels, and claim the space as long as it is possible to change allocation decisions. Overruling is governed by certain priority calculations.

3 Experimental Results

We build a simulation model to analyze the performance of the algorithm given different operating conditions and different infeed lengths. We model a merge configuration with four infeeds. Our main input variables are: parcels' lengths, parcels' inter-arrival times, and the density of flying through parcels. The latter is described in terms of the space occupancy on the merge conveyor, when appearing at the *fly through point*, which is the upstream boundary of the merge configuration (see Figure 1). Normally, parcels are loaded successively when they belong to the same batch, but it takes some time until a new batch starts loading. In the simulation experiments the distribution of inter-arrival times within the same batch is an input variable, using three different uniform distributions. However, batch arrivals are always embedded in the experiments with a certain probability in the same manner, and so provide an input that is not varied. We first perform simulation experiments to tune some parametric values used in the algorithm. Later, we report on our main output measures, which reflect the aforementioned objectives: First, space utilization on the merge conveyor. Second, relative difference in waiting times, calculated as the difference between the maximum and minimum values of total waiting times on the modeled infeeds, divided by the maximum value of total waiting time among infeeds. Moreover, we report on the frequency of reallocations.

In standard operating conditions, the algorithm achieves good results (utilization exceeds 90%, and relative difference in waiting times is 1-2%). Main remarks are: First, as inter-arrival times increase, space utilization on the merge conveyor drops. The explanation of this trend is intuitive. When inter-arrival times are long, much space on the merge conveyor is missed as the conveyor moves. Second, given a fixed range of inter-arrival times, increasing the density of flying through parcels increases utilization of the merge conveyor in most of the cases. As the density of flying through parcels increases, parcels on the infeeds are more likely to wait for merge spaces, and so more time is available to make more informed allocation decisions. Allocation decisions are more likely to be made by Process B (see Section 2), which considers space utilization in calculating priorities. Third, for short and medium ranges of inter-arrival times, varying the density of flying through parcels has a negligible effect (1%) on the relative difference in waiting times. Fourth, an operating condition where long inter-arrival times are combined with no flying through parcels, results in relatively high imbalance measures (20%). In this case, the effect of process B is limited, because the possibility that parcels from different infeeds are simultaneously waiting for space allocations is small. Moreover, the effect of process A is limited, because changing space allocation decisions becomes less feasible. The reason is late awareness of new incoming parcels, due to long inter-arrival times, and less waiting times of available parcels. Less waiting time is caused by no or low density of flying through parcels. We discuss that this condition may not create a problem in practice. We also argue that better workload balancing may not be achievable even with other algorithms, as long as the layout of the merge configuration is not altered. However, we show how small variations in the lengths of the infeeds helps overcome this issue. Furthermore, we experiment with different

infeed lengths, the impact of the reallocation process, and the relation between the number of reallocations executed and the length of the infeeds. We also report on the distribution of waiting time, and of reallocations, among infeeds.

4 Conclusion

The generic PBA works on maximizing throughput as the main objective, and then executes a reallocation procedure to balance the workload as the second objective while not deteriorating throughput. Simulation shows that the algorithm gives good results for different layouts of the merge configuration, and under varying operating conditions. An important lesson learned is that a layout design with certain control rules can have dramatic results. However, an integrated solution of layout design and control rules may result in significant improvements. This study was conducted in cooperation with an international supplier of express parcel sorting systems, where the algorithm is currently being implemented (Company name is concealed for confidentiality reasons). Going back to the wide picture of our research, which is planning and control of AMHSs, we see this study as an important component of the overall project. We provide a generic algorithm that can be adapted to different operational conditions while maintaining the same generic structure. For example, the calculations of priorities within the algorithm may include different criteria. Although the merge operation occurs in AMHSs of different sectors, it is mostly critical for express parcel sorting systems. This is due to the emphasis on providing sorting systems with high throughput, and so it is crucial to use the space within the system as efficiently as possible. Since the generic PBA can be applied to different merge configurations, it can replace the various customized control algorithms that are currently used for different systems and different sectors.

References

1. Chen F.F. Sturges R.H. Babiceanu, R.F. Framework for the control of automated material-handling systems using the holonic manufacturing approach. *International Journal of Production Research*, 42(17): 3551–3564, 2004.
2. Boyd N.P. Whorms H.H. Dilts, D.M. The evolution of control architectures for automated manufacturing systems. *Journal of Manufacturing Systems*, 10(1): 79–93, 1991.
3. Bozer Y.A. Hsieh, Y.-J. Analytical modeling of closed-loop conveyors with load recirculation. volume 3483, pages 437–447, 2005.
4. Douglas L. McWilliams. A dynamic load-balancing scheme for the parcel hub-scheduling problem. *Computers & Industrial Engineering*, 57(3): 958–962, 2009.
5. Douglas L. McWilliams. Iterative improvement to solve the parcel hub scheduling problem. *Computers & Industrial Engineering*, 59(1): 136–144, 2010.
6. El-Nashar A. Nazzal, D. Survey of research in modeling conveyor-based automated material handling systems in wafer fabs. 2007.
7. S.E. Shladover. Operation of merge junctions in a dynamically entrained automated guideway transit system. *Transportation Research Part A: General*, 14(2): 85–112, 1980.

II.2 Discrete Optimization, Graphs & Networks

Chair: Prof. Dr. Sven Oliver Krumke (Universität Kaiserslautern)

Discrete Optimization is a branch of applied mathematics which has its roots in Combinatorics and Graph Theory. Discrete Optimization is located on the boundary between mathematics and computer science where it touches algorithm theory and computational complexity. It has important applications in several fields, including artificial intelligence, mathematics, and software engineering. As opposed to continuous optimization, the variables used in the mathematical program (or some of them) are restricted to assume only discrete values, such as the integers. Two notable branches of discrete optimization are: combinatorial optimization,which refers to problems on graphs, matroids and other discrete structures, and integer programming.

These branches are closely intertwined however since many combinatorial optimization problems can be modeled as integer programs and, conversely, integer programs can often be given a combinatorial interpretation. Graphs and networks play an important role in modelling problems; they appear naturally in various applications and there is a variety of efficient algorithms available for solving "standard" problems on networks. However, many problems in Discrete Optimization are "difficult" to solve (for instance NP-hard). Thus, especially for problems of practical interest one would like to compromise on the quality of the solution for the sake of computing a suboptimal solution quickly.

Securely Connected Facility Location in Metric Graphs

Maren Martens and Andreas Bley

Abstract Connected facility location problems arise in many different applications areas, such as transportation, logistics, or telecommunications. Given a set of clients and potential facilities, one has to construct a connected facility network and attach the remaining clients to the chosen facilities via access links. Here, we consider interconnected facility location problems, where we request 1- or 2-connectivity in the subnetwork induced by the chosen facilities alone, disregarding client nodes. This type of connectivity is required in telecommunication networks, for example, where facilities represent core nodes that communicate directly with each other. We show that the interconnected facility location problem is strongly NP-hard for both 1- and 2-connectivity among the facilities, even for metric edge costs. We present simple randomized approximation algorithms with expected approximation ratios 4.40 for 1-connectivity and 4.42 for 2-connectivity. For the classical 2-connected facility location problem, which allows to use non-facility nodes to connect facilities, we obtain an algorithm with expected approximation guarantee of 4.26.

1 Introduction

The desire to cost-efficiently plan large networks where a set of clients is served by several interacting facilities has given rise to a spate of research on combined facility location and connectivity problems. Such problems include, e.g., the design of optical telecommunication networks where clients are (directly) attached to a connected core network. While most investigations on connected facility problems allow the core network to contain clients as Steiner nodes, this is prohibited in most network design problems in telecommunications; see, e.g., [9].

Maren Martens
Axxom Software AG, P.-Gerhardt-Allee 46, 81245 München, e-mail: maren.martens@axxom.com

Andreas Bley
TU Berlin, Straße des 17. Juni 136, 10623 Berlin, e-mail: bley@math.tu-berlin.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_46,

We consider *inter*connected facility location problems, where the core network induced by the open facilities must be connected. In order to ensure network reliability, which is a big issue in telecommunications, it is common to require this core network to be 2-connected. For all problems considered in this paper, we are given a complete undirected graph $G = (V,E)$, a set $P \subseteq V$ of potential facilities, nonnegative opening costs $f: P \to \mathbb{R}_{\geq 0}$ and nonnegative edge costs $c: E \to \mathbb{R}_{\geq 0}$ serving the triangle inequality. For simplicity, we let $n = |V|$ and $m = |E|$. The *interconnected facility location problem (iCFL)* asks for a set $\mathcal{F} \subseteq P$ of facilities to be opened, for edges $E_C \subseteq \delta(\mathcal{F}) := \{uv \in E \mid u \notin \mathcal{F}, v \in \mathcal{F}\}$ that directly connect every $u \notin \mathcal{F}$ to some $v \in \mathcal{F}$, and for edges $E_F \subseteq \{uv \in E \mid u,v \in \mathcal{F}\}$ such that $(\mathcal{F},E_F)$ is a tree. We call $(\mathcal{F},E_F)$ the *core network* of the solution. In the *2-interconnected facility location problem (2-iCFL)* E_F must be chosen such that $(\mathcal{F},E_F)$ is 2-node-connected. Even for metric edge costs, connection cost can be reduced by using nodes not in $\mathcal{F}$ as Steiner nodes in the core. In the corresponding *2-connected facility location (2-CFL)* we ask for a set $\mathcal{F} \subseteq P$ of facilities to open and edge sets $E_C \subseteq \delta(\mathcal{F})$ and $E_F \subseteq E$, such that (V,E_F) contains two node-disjoint paths between every pair of nodes in $\mathcal{F}$ and E_C contains an edge uv with $v \in \mathcal{F}$ for every node $u \notin \mathcal{F}$. Here, the two node-disjoint paths in E_F connecting two open facilities may use non-facility nodes. In all problems the objective is to minimize $\sum_{v\in\mathcal{F}} f_v + \sum_{e\in E_C} c_e + M \cdot \sum_{e\in E_F} c_e$, where $M \geq 1$ reflects possibly higher costs for interfacility connections. We assume that $M/|V| < \varepsilon$.

We restrict our attention to 2-node-connectivity as, in the case of metric edge costs, 2-node- and 2-edge-connectivity coincide. The following procedure turns a 2-edge-connected graph G into a 2-node-connected graph without increasing cost:

ECON-2-NCON(Frederickson Ja'Ja'[5]): While there is a node v whose removal would disconnect G: Find edges uv and vw where u and w are in different maximal 2-node-connected subgraphs of G. Then replace uv and vw by uw.

Connected facility location (CFL) and variants thereof, which allow non-facilities as Steiner nodes in the core network and paths instead of direct edges as connections from clients to facilities, have been widely studied. Eisenbrand et al.[4] gave the currently best approximation algorithms for CFL, k-CFL, and tour-CFL, obtaining expected approximation ratios of 4, 6.85, and 4.12, respectively. Problems that implicitly require 2-connectivity among open facilities (excluding clients) have been studied as *ring star*, *tour-CFL*, or *hub location* problems. In these problems, the core network has to form a ring or a complete graph. ILP formulations as well as approximation algorithms have been investigated for these special, restricted versions of connected facility location problems, see [6, 7, 9, 10] for example. The more general 2-iCFL was studied in [2], where it was shown that it is NP-hard to approximate 2-iCFL within a factor $c \log n$, for some constant $c > 0$, and cut based ILP formulations that allow to solve instances of several hundreds of nodes have been presented.

In this paper we prove strong NP-hardness and present randomized approximation algorithms for all three problems iCFL, 2-iCFL, and 2-CFL in graphs with metric edge costs.

2 Complexity

To show that 2-CFL and 2-iCFL are strongly NP-hard, we present a reduction from the HAMILTONIAN CYCLE problem: Given an undirected graph $H = (V_H, E_H)$, we have to decide if H contains a cycle that visits every node exactly once.

Theorem 1 *2-CFL and 2-iCFL are strongly NP-hard in graphs with metric costs.*

Proof. Let $H = (V_H, E_H)$ be the given instance of HAMILTONIAN CYCLE. Consider the complete graph $G = (V, E)$ on the $2|V_H|$ nodes $V = \{v, v' \mid v \in V_H\}$ with edge costs $c_e = 1$ for all $e = E_H \cup \{vv' \mid v \in V_H\}$ and $c_e = 2$ otherwise. Note that these costs are metric. We set $M = 1$ and $P = V_H$. Facilities can be opened at no charge.

One easily verifies that, for both problems 2-CFL and 2-iCFL, solutions that open all facilities in V_H, connect them via a Hamiltonian cycle along V_H (if any exists) and use the edges of cost 1 to connect V' to V_H are the only solutions of cost (at most) $2n$. All other solutions for 2-CFL or 2-iCFL have cost at least $2n+1$. Thus, we have a solution of cost (at most) $2n$ iff there is a Hamiltonian cycle in H. □

For iCFL, we use a reduction from MINIMUM SET COVER: Given a collection C of subsets of a finite set S, find subset $C' \subseteq C$ of minimum cardinality such that every element in S belongs to at least one member of C'.

Theorem 2 *iCFL is strongly NP-hard in graphs with metric costs.*

Proof. From an instance of MINIMUM SET COVER we construct the following instance for metric iCFL: Let V contain one node v_c, for every $c \in C$, one node v_s, for every $s \in S$, and one extra node v. For edges vv_c and $v_c v_s$ with $s \in c \in C$, the cost is 1, all other edges have cost 2. We set $M = 1$ and $P = \{v_c \mid c \in C\} \cup \{v\}$ with opening cost 1, for v_c with $c \in C$, and opening cost 0 for v.

Note that the cost of any iCFL solution opening k facilities from $P \setminus \{v\}$ is at least $|C| + |S|$ for the connectivity plus k for opening facilities. Moreover, a total cost of exactly $|C| + |S| + k$ is achieved iff only edges of cost 1 are chosen in the solution.

Let C^* be a minimum set cover. Then opening all facilities v_c with $c \in C^*$ plus v and using all edges vv_c with $c \in C$ plus one edge, for every $s \in S$, which connects s to some $c \in C^*$ with $s \in c$, yields a feasible iCFL solution of cost $K := |C| + |S| + |C^*|$. On the other hand, every iCFL solution of cost at most K provides a minimum set cover: If the considered solution opens $|C^*|$ facilities, its cost is at most K only if all chosen edges have cost 1. Therefore, every node v_s must be connected to an open facility v_c with $s \in c$ and so the set of open facilities yields a minimum set cover. If, however, $k < |C^*|$ facilities are opened, there is at least one v_s that is connected to an open facility via an edge of cost 2. For this v_s, we delete its connecting edge and in return open a facility at some node v_c with $s \in c$ and connect v_s to this new facility. This does not increase cost. We continue opening facilities and reconnecting clients, until all v_s's are connected to open facilities via edges of cost 1. Eventually, (at most) $|C^*|$ facilities are open, since the cost of the final solution is still (at most) K and–by the above arguments–any solution with more than $|C^*|$ open facilities costs more than K. Thus, iCFL has a solution of cost (at most) $|C| + |S| + k$ iff MINIMUM SET COVER has a set cover of size at most k. □

3 Approximating iCFL

In the following, we use the techniques introduced in [4] to construct an approximation algorithm for the metric iCFL problem. The algorithm uses some parameter $\alpha \in (0,1]$ that will be fixed at the end of this section. The pair $(\mathcal{F}_U, \sigma_U)$ denotes the solution for the (unconnected) facility location problem defined by our input data and solved in the first step of the algorithm. There, $\mathcal{F}_U$ is the set of open facilities and σ_U is a map that assigns every client to an open facility. For simplicity, we let σ_U assign open facilities to themselves.

Algorithm 1 for general metric iCFL is the following:

1. Compute an approximate solution $U = (\mathcal{F}_U, \sigma_U)$ for the (unconnected) facility location problem.
2. Choose one node in V uniformly at random and mark it. Mark every other node independently with probability α/M. Let D be the set of marked nodes.
3. Open facility $i \in \mathcal{F}_U$ if there is at least one marked node in $\sigma_U^{-1}(i)$. Let $\mathcal{F}$ be the set of open facilities.
4. Compute a MST $\mathcal{T}$ on the subgraph induced by $\mathcal{F}$.
5. Return $A = (\mathcal{F}, \mathcal{T}, \sigma)$ where σ assigns each $v \in V$ to a closest open facility in $\mathcal{F}$.

Let O_U and C_U be the opening and connection cost in U. Further let $OPT = (\mathcal{F}^*, \mathcal{T}^*, \sigma^*)$ be an optimal solution for iCFL where $\mathcal{F}^*$ is the set of open facilities, $\mathcal{T}^*$ is a minimum spanning tree among those, and σ^* is a cost minimal client-to-facility map. The opening, client connection, and core connection cost for A and OPT are denoted as O, C, T and O^*, C^*, T^*, respectively. Here, T and T^* denote the core connection cost including the multiplication with M, i.e., $T = M \cdot \sum_{e \in \mathcal{T}} c(e)$ and $T^* = M \cdot \sum_{e \in \mathcal{T}^*} c(e)$. Obviously, $O \le O_U$. For now, we keep this as an upper bound for O and continue with T. Later on, we bound C_U and O_U by C^* and O^*.

Lemma 1 $E[T] \le 2T^* + 2(\alpha+\varepsilon)(C^* + C_U)$

Proof. The following gives us a spanning tree on $\mathcal{F}$: Take $\mathcal{T}^*$ and augment it in two steps: First add an edge between every client $v \in D$ and $\sigma^*(v)$, then add an edge between every $f \in \mathcal{F} \setminus (D \cup \mathcal{F}^*)$ and some $v \in \sigma_U^{-1}(f) \cap D$. In the resulting graph, which contains $\mathcal{F}$, double each edge in order to make the graph Eulerian. Then take a Euler path with shortcuts around any nodes not in $\mathcal{F}$ or those already visited. The resulting path on $\mathcal{F}$ has expected cost at most

$$
\begin{aligned}
&2 \sum_{e \in \mathcal{T}^*} c(e) + 2 \sum_{v \in V} \left(\frac{\alpha}{M} + \frac{1}{n} \right) c(v, \sigma^*(v)) + 2 \sum_{v \in V} \left(\frac{\alpha}{M} + \frac{1}{n} \right) c(v, \sigma_U(v)) \\
&= \frac{2}{M} T^* + 2 \left(\frac{\alpha}{M} + \frac{1}{n} \right) C^* + 2 \left(\frac{\alpha}{M} + \frac{1}{n} \right) C_U \qquad .
\end{aligned}
$$

As $\mathcal{T}$ costs no more than the path constructed above and, by assumption, $M/n \le \varepsilon$, we have $E[T] \le M((2/M)T^* + 2(\alpha/M + 1/n)(C^* + C_U)) \le 2T^* + 2(\alpha+\varepsilon)(C^* + C_U)$. □

For $E[C]$, we use the same bound as derived in [4] for CFL. This is $E[C] \le 2C^* + C_U + (0.807/\alpha)T^*$. To solve the facility location problem in the first step, we use the

bifactor approximation algorithm by Byrka and Aardal [1], which returns a solution U with facility cost $O_U \leq \delta O^*$ and connection cost $C_U \leq (1+2e^{-\delta})C^*$ for any $\delta \geq 1.678$. With a proper choice of α and δ, we obtain the following result.

Theorem 3 *Algorithm 1 is an expected 4.40-approximation algorithm for iCFL.*

Proof. Adding the bounds derived above yields $E[O+T+C] \leq O_U + 2T^* + 2(\alpha+\varepsilon)(C^*+C_U) + 2C^* + C_U + (0.807/\alpha)T^*$. With the bifactor algorithm's approximation quarantees, this implies $E[O+T+C] \leq (2(\alpha+\varepsilon)+2+(2(\alpha+\varepsilon)+1)(1+2e^{-\delta}))C^* + \delta O^* + (2+0.807/\alpha)T^*$. For ε sufficiently small, $\alpha = 0.3397$, and $\delta = 4.40$, we obtain $E[O+T+C] \leq 4.40(O^*+T^*+C^*)$. □

4 Approximating 2-CFL and 2-iCFL

Changing Step 4 of Algorithm 1, we now derive approximation algorithms for metric 2-CFL and 2-iCFL. Algorithm 2 for metric 2-CFL is obtained by replacing Step 4 in Algorithm 1 by:

4a. Compute a ρ_{2CS}-approximate 2-connected subgraph on D. Augment this graph by adding two parallel edges between each client $v \in D$ and its facility $\sigma_U(v) \in \mathcal{F}$. Apply ECON-2-NCON to turn this graph into a 2-node-connected graph $\mathcal{S}$.

Theorem 4 *Algorithm 2 is an expected 4.26-approximation algorithm for 2-CFL.*

Proof. Let $OPT = (\mathcal{F}^*, \mathcal{S}^*, \sigma^*)$ be an optimal solution of the 2-CFL problem, where $\mathcal{S}^*$ is a minimum 2-connected subgraph among the open facilities. As before, O_U and C_U denote the opening and connection cost in U, while O, C, S and O^*, C^*, S^* denote the opening, client connection, and core connection cost of the solution of Algorithm 2 and of the optimal solution, respectively. Clearly, $O \leq O_U$.

The following yields a 2-connected subgraph on D: Take $\mathcal{S}^*$, add two edges between every client in D and its closest node in $\mathcal{S}^*$, and apply ECON-2-NCON to make the graph 2-node-connected. The expected cost of this subgraph is at most $\sum_{e\in\mathcal{S}^*} c(e) + 2\sum_{v\in V\setminus\mathcal{F}^*}(\alpha/M+1/n)c(v,\sigma^*(v)) = (1/M)S^* + 2(\alpha/M+1/n)C^*$. Thus, the expected cost of the subgraph $\mathcal{S}$ computed in Step 4a can be bounded by $(\rho_{2CS}/M)S^* + 2\rho_{2CS}(\alpha/M+1/n)C^*$. The expected cost of adding the edges in Step 4a is $(\alpha/M+1/n)C_U$ and thus $E[S] \leq \rho_{2CS}(S^*+2(\alpha+\varepsilon)C^*) + 2(\alpha+\varepsilon)C_U$.

The connection cost can be bounded using the core connection game introduced in [4]. There it is proven that, if the core S^* is a cycle, the expected connection cost will not exceed $2C^* + C_U + S^*/(2\alpha)$. It is further shown, how this result can be adjusted to other core graph structures. Using a result by Monma and Munson [8], we know that we can turn our optimal core graph S^* into a cycle by losing a factor at most $4/3$. So we can prove that $E[C] \leq 2C^* + C_U + 2S^*/(3\alpha)$.

As in Theorem 3, we have $C_U \leq (1+2e^{-\delta})C^*$ and $O_U \leq \delta O^*$ for any $\delta \geq 1.678$. Thus, $E[O+S+C] \leq (\rho_{2CS}+2/(3\alpha))S^* + \delta O^* + (2\rho_{2CS}(\alpha+\varepsilon)+2+(1+2(\alpha+\varepsilon))(1+2e^{-\delta}))C^*$. For ε sufficiently small, $\alpha = 0.2436$, $\delta = 4.26$, and $\rho_{2CS} = 1.5$ (see [5]), we obtain $E[O+S+C] \leq 4.26(O^*+S^*+C^*)$. □

To solve the metric 2-iCFL problem, an obvious idea would be to compute a minimum 2-connected subgraph on $\mathcal{F}$ in Step 4. However, since best know approximation ratio for the metric 2-connected subgraph problem is not smaller than that for the metric TSP, it eventually turns out to be better to solve tour-CFL instead. The corresponding Algorithm 3 is obtained by replacing Step 4 in Algorithm 1 by:

4b.Compute a ρ_{TSP}-approximate TSP-tour on D. Augment this tour by adding two shortest paths between each client $v \in D$ and its facility $\sigma_U(v) \in \mathcal{F}$. Eventually, compute a Euler tour on this graph and shortcut it to a TSP-tour $\mathcal{T}$ on $\mathcal{F}$.

Theorem 5 *Algorithm 3 is an expected 4.42-approximation algorithm for 2-iCFL.*

Proof. Consider tour-$OPT = (\mathcal{F}^*, \mathcal{T}^*, \sigma^*)$ where $\mathcal{T}^*$ is an optimal TSP-tour on $\mathcal{F}^*$ of cost T^*. With the same notation as above, we have $O \le O_U$ and $T^* \le (4/3)S^*$. With the arguments in [4] it can be shown that $E[S] \le \rho_{\text{TSP}}(T^* + 2(\alpha+\varepsilon)C^*) + 2(\alpha+\varepsilon)C_U$ and $E[C] \le 2C^* + C_U + T^*/(2\alpha)$. Thus, $E[S] \le \rho_{\text{TSP}}(\frac{4}{3}S^* + 2(\alpha+\varepsilon)C^*) + 2(\alpha+\varepsilon)C_U$ and $E[C] \le 2C^* + C_U + 2S^*/(3\alpha)$. Using these bounds, we get $E[O+S+C] \le ((4/3)\rho_{\text{TSP}} + 2/(3\alpha))S^* + \delta O^* + (2\rho_{\text{TSP}}(\alpha+\varepsilon) + 2 + (1+2(\alpha+\varepsilon))(1+2e^{-\delta}))C^*$. With ε sufficiently small, $\alpha = 0.2765$, $\delta = 4.42$, and $\rho_{\text{TSP}} = 1.5$ (see [3]) the theorem follows. □

Acknowledgements This works was supported by the Federal Ministry of Education and Research (BMBF), Germany, as part of the EUREKA project "100GET Technologies" and by the DFG Research Center MATHEON–"Mathematics for key technologies".

References

1. J. Byrka and K. Aardal. An optimal bifactor approximation algorithm for the metric uncapacitated facility location problem. *SIAM Journal on Computing*, 2010. to appear.
2. M. Chimani, M. Kandyba, and M. Martens. 2-interconnected facility location: Specifications, complexity results, and exact solutions. Technical Report TR09-1-008, Computer Science Department, Technical University Dortmund, 2009.
3. N. Christofides. Worst-case analysis of a new heuristic for the traveling salesman problem. Technical Report 388, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, 1976.
4. F. Eisenbrand, F. Grandoni, T. Rothvoß, and G. Schäfer. Approximating connected facility location problems via random facility sampling and core detouring. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1174–1183, 2008.
5. G. N. Frederickson and J. Ja'Ja'. On the relationship between the binconnectivity augmentation and traveling salesman problems. *Theoretical Computer Science*, 19: 189–201, 1982.
6. S. Kedad-Sidhoum and V. H. Nguyen. An exact algorithm for solving the ring star problem. http://www.optimization-online.org/DB_HTML/2008/03/1942.html, March 2008.
7. M. Labbe, G. Laporte, I. Rodriguez Martin, and J. J. Salazar Gonzalez. The ring star problem: Polyhedral analysis and exact algorithm. *Networks*, 43(3): 177–189, 2004.
8. C. L. Monma, B. S. Munson, and W. R. Pulleyblank. Minimum-weight two-connected spanning networks. *Mathematical Programming*, 46(2): 153–171, 1990.
9. M. Pioro and D. Medhi. *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann Publishers, 2004.
10. R. Ravi and F. S. Salman. Approximation algorithms for the traveling purchaser problem and its variants in network design. In *Proceedings of the 7th Annual European Symposium on Algorithms*, pages 29–40, 1999.

Network Flow Optimization with Minimum Quantities

Hans Georg Seedig

Abstract Economies of scale often appear as decreasing marginal costs, but may also occur as a constraint that at least a minimum quantity should be produced or nothing at all. Taking the latter into account can support decisions about keeping a production site, setting up a new one, establishing new routes in logistics network, and many more. In our work, we define the corresponding problem MINIMUM COST NETWORK FLOW WITH MINIMUM QUANTITIES and prove its computational hardness. Following that, we devise a tailored Branch-&-Bound algorithm with an efficient update step that usually affects only small parts of the network. Experiments are conducted on real problems and artificial networks. In our results, the update step made the inevitable but well-studied subproblem of the initial computation of a minimum cost network flow problem taking most of the actual total running time. As our method does also compare favorably with a heuristic algorithm that has been implemented before, we recommend it for practical use.

1 Introduction

The problem we study is derived form the classical MINIMUM COST NETWORK FLOW Problem (MCNF). Because of their immanent applicability to cost-intensive tasks, MCNF and many of its variants have been studies extensively in the past. Among these are the problems where the costs per unit amount are increasing or decreasing (convex or concave costs) and where the amount routed through a network changes on the way (generalized flow problem). Treating other than fixed costs per unit amount often reflects the real cost structure of a given real-world problem.

Hans Georg Seedig
Department of Informatics
Technische Universität München
85748 Garching bei München, Germany
e-mail: seedigh@in.tum.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_47,

Costs amortize with bigger amounts, for example at production sites. Having a factory building a single car makes this car a lot more expensive compared to a factory where hundreds of cars are built each day. On the other side, the marginal costs might rise for for larger numbers. If the facilities are used to capacity, an increase in output demands for the construction of new facilities, additional work force and so on. These effects are often referred to as *economy of scale*.

A new question was raised recently where closing some production facilities was considered favorably for lowering costs. While the fixed setup costs that occur as soon as anything is produced at all could not be quantified, the customer named a threshold amount of output from which on a location was considered worth being kept. This amount is what we call *Minimum Quantity* (or *MQ*).

So, the problem to solve deals only with constant marginal costs but restricts the solution space as suggestions where a location had an output level above zero but below the MQ would be considered ineligible.

2 Preliminaries

In this work, a *network* $N = (V, E, u, c)$ is a digraph (V, E) equipped with a *capacity* function $u : E \to \mathbb{R}^+ \cup \{\infty\}$ and a *cost* function $c : E \to \mathbb{R}$.[1] At the nodes, a *demand* function $b : V \to \mathbb{R} \cup \{-\infty\}$ is given. We call each node s with $b(s) > 0$ a *source*, having *supply*, and every node t with $b(t) < 0$ a *sink*. As every network with more than one source and/or more than one sink is easily transformed to an equivalent one with a single source and a single sink, we assume the latter to be the case. Let $|V| =: n$ and $|E| =: m$.

A *feasible flow* in a network N is a function $f : E \to \mathbb{R}^+$ satisfying the *capacity constraints* for each e in E: $f(e) \leq u(e)$ and flow conservation constraints

$$\sum_{e_1=(v',v)} f(e_1) - \sum_{e_2=(v,v'')} f(e_2) = b(v) \quad \forall v \in V. \tag{1}$$

Given a network N, we define the *flow network* $N_F(f) = (V, E^f, u^f, c^f)$ where E^f is the set of all edges from N having a positive flow value in f, the capacity u^f equals f for all these edges and the costs on these edges are the same as in N.

Similarly, the *residual network* $N_R(f) = (V, E_R^f, u_R^f)$ with costs c_R^f is the network with capacities indicating how much flow can still be sent at which price along an edge without violating the capacity bound and how much can be sent back to a node without the flow being negative while reducing the total costs.

[1] We will use vector notation u_e and c_e to denote the capacity and cost of edge e.

3 Minimum Cost Network Flow

The central problem regarding network flows is that of finding a feasible flow with minimum cost that satisfies all demands. For a network with one source s,one sink t, and demand d, the problem is defined as follows.

MINIMUM COST NETWORK FLOW (MCNF)
Input: Network $N = (V,E)$, $s,t \in V$, capacities u_e and costs c_e for all $e \in E$ and demands per node $b_t = -b_s = d, b_v = 0 \forall v \in V \setminus \{s,t\}$
Task: Find a feasible flow $f = (f_1, \ldots, f_m)$ that minimizes

$$\sum_{e \in E} c_e \cdot f_e.$$

Different types of algorithms have been devised for MCNF with more than one allowing a strongly polynomial running time. Today, the most popular methods (see, e.g., [7, 4]) use a scaling approach of one kind or another. An interesting optimality criterion for an optimal flow f^* is that the residual network $N_R(f^*)$ contains no cycles of total negative cost. This implies an algorithm that successively identifies and eliminates such cycles by increasing the flow along the cycle's edges. When in each step the cycle with minimum mean costs is canceled, a strongly polynomial bound can be achieved [5].

4 Minimum Cost Network Flow with Minimum Quantities

We define the problem of a minimum cost network flow with minimum quantities on the first layer or MCNFMQ as a mixed integer linear program (MILP).

MCNF WITH MINIMUM QUANTITIES (MCNFMQ)
Input: Network $N = (V,E)$ with $s,t \in V$, capacities $u_e \forall\, e \in E$, a demand $d \in \mathbb{N}$ and minimum quantity lots $\lambda_e \neq 0$ on edges $(s,v) \in E$.
Task: Find $x = (x_1, \ldots, x_m)$ and a feasible flow $f = (f_1, \ldots, f_m)$ that satisfies

$$\sum_{e=(s,u) \in E} f_e = \sum_{e=(w,t) \in E} f_e = val(f) = d,$$

$$\lambda_e x_e \leq f_e \leq x_e u_e \qquad \forall\, e \in E$$

and minimizes

$$\sum_{e \in E} c_e \cdot f_e.$$

4.1 Complexity of MCNFMQ

To show the NP-completeness of MCNFMQ, we use the known completeness of the problem SUBSETSUM for the reduction. SUBSETSUM is a variant of the KNAPSACK problem and is indeed NP-complete as shown by [3].

SUBSETSUM
Input: $A = \{a_1, \ldots, a_n\}, a_i \in \mathbb{N}$
$k \in \mathbb{N}$
Question: Does there exist an $I \subset \{1, \ldots, n\}$ such that $\sum_{i \in I} a_i = k$?

Theorem 1 *MCNFMQ is NP-complete.*

Proof. Given an instance of SUBSETSUM*, define an instance of* MCNFMQ *as follows.*

- $V = \{s, t, t_0, v_1, \ldots, v_n\}$ *nodes*
- $E \times \mathbb{N}^2 = \{(s, v_1, a_1, a_1), \ldots, (s, v_n, a_1, a_n)\} \cup \{(v_1, t_0, 0, \infty), \ldots, (v_n, t_0, 0, \infty)\} \cup \{(t_0, t, 0, k)\}$ *edges with MQ lot and upper capacity.*

For the network $N = (V, E)$ with MQ lots and upper capacities as defined and all $c_e = 0$, it is straightforward to check that the given instance of SUBSETSUM *is a* YES *instance if and only if there is a feasible flow in the corresponding MCNFMQ.*

4.2 Algorithm for MCNFMQ

Because of the computational hardness of MCNFMQ, it is justified to look at possible super-polynomial algorithms. First, we introduce the notion of a *configuration* which is a tuple of decision variables, one for each MQ-constrained edge, forcing the flow on this edge to be zero or above the MQ lot. Each such tuple defines a new problem that turns out to have an MCNF equivalent. Our algorithm is a Branch-&-Bound method on this set of configurations.

The algorithm works as follows. We start with a completely relaxed MCNFMQ (which is an MCNF) and efficiently compute its solution. On edges where the solution violates the MQ constraint, we branch to obtain two new configuration. After applying Wagner's transformation [8] we have corresponding MCNFs that can be solved efficiently with the methods discussed earlier.

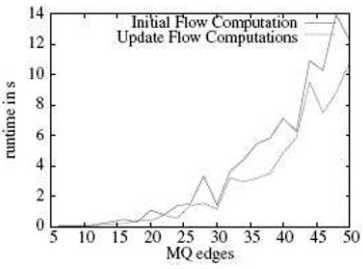
We increased the performance of our algorithm dramatically by implementing an update step, using the previously computed solution of the parent configuration for the new setting instead of computing the solution to the new MCNFs from scratch. It is a rerouting procedure, where the flow is increased (decreased) along shortest (longest) s-t-paths in the residual (flow) network to first match the new MQ constraint, followed by the elimination of cycles with total negative cost in the residual network to regain optimality, and decrease (increase) the flow again along longest (shortest) s-t-paths in the flow (residual) network to reach feasibility.

5 Experimental Results

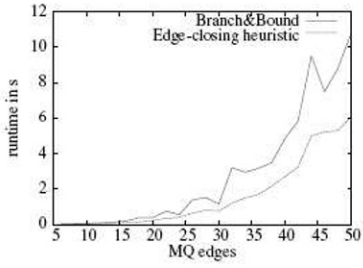
We compared our Branch-&-Bound (B&B) method with an established heuristic that has been used in the past. That method basically also starts with a relaxed solution, but then only closes edges with violated MQ-constraints and does no backtracking. As soon as a feasible solution is found, it is returned but neither its optimality nor the discovery of any existing feasible solution is guaranteed.

For our experiments, we had a data set of an international company to our avail and did also generate similarly-structured artificial networks. These are basically a fully connected graph where each of the vertices is to be thought of as a depot and is connected to the source and the sink. The MQ-constrained edges are the ones connected to the source and are all assigned the same MQ lot. Costs and capacities of all edges were randomly chosen and above the MQ lot where applicable.

Our findings are that the average optimality gap of the rather simple heuristic is about 4% for our models, tested against the B&B method that is guaranteed to compute the exact solution. The update procedure proved to be effective as on average the time needed for all updates during the exploration of the B&B tree roughly equaled the time needed for the inevitable computation of the very first relaxed solution as can be seen Figure 1(a). With this, the B&B method takes only slightly more time than the heuristic in our experiments as depicted in Figure 1(b) while usually computing solutions for about 3 times as many configurations during its execution.



(a) Execution time of B&B split into computation of initial solution and update phase



(b) Comparison of B&B and edge-closing heuristic considering execution time of the update phase

Fig. 1 Execution time comparisons on randomly generated networks with two models, averaged over 30 runs.

6 Conclusions and Outlook

We introduced a new type of constraints in the context of Minimum Cost Network Flow that has obvious applications to real world problems. We were able to prove the hardness of the general MCNFMQ problem and subsequently devised a tailored Branch-&-Bound algorithm with an update method that showed to be very efficient in practice. That way, the computational complexity of real world instances was basically reduced to the well-studied classical MCNF problem for which very fast algorithms have already been implemented [4]. The same is true for the subproblem of identifying cycles of negative length in the residual network we used for the update step [1]. This makes our approach suitable for actual problems in logistics.

Possible future work is to investigate more on different cost models to account for economies of scale or other effects. There has been some work on complexity and algorithms for network flow problem with nonlinear cost functions (,e.g., [6], [2]) but we think that this is worth extending. While the resulting problems are mostly hard, it would still be interesting to devise tailored algorithms for practical use.

We would be interested in seeing our approach extended to multi-commodity problems where the constraints are coupled. It is also a remaining question wether the hardness of MCNFMQ holds if all possible minimum quantity lots are known to be the same.

Acknowledgements The author thanks Prof. Dr. A. Taraz and Dr. Mirko Eickhoff for helpful discussions. Part of this work was done while the author was at Axxom Software AG, Munich, Germany.

References

1. Boris Vasilievich Cherkassky and Andrew Vladislav Goldberg. Negative-cycle detection algorithms. *Mathematical Programming*, 85: 277–311, 1999.
2. Dalila Benedita Machado Martins Fontes, Eleni Hadjiconstantinou, and Nicos Christofides. A branch-and-bound algorithm for concave network flow problems. *Journal of Global Optimization*, 34(1): 127–155, 2006.
3. Michael Randolph Garey and David Stifler Johnson. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. W. H. Freeman, January 1979.
4. Andrew Vladislav Goldberg. An efficient implementation of a scaling minimum-cost flow algorithm. *Journal of Algorithms*, 22: 1–29, 1997.
5. Andrew Vladislav Goldberg and Robert Endre Tarjan. Finding minimum-cost circulations by canceling negative cycles. *Journal of the ACM*, 36(4): 873–886, 1989.
6. Dorit Simona Hochbaum. Complexity and algorithms for nonlinear optimization problems. *Annals of Operations Research*, 153(1): 257–296, 2007.
7. James Berger Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations Research*, 41(2): 338–350, 1993.
8. Harvey Maurice Wagner. On a class of capacitated transportation problems. *Management Science*, 5(3): 304–318, 1959.

A Hybrid Genetic Algorithm for Constrained Combinatorial Problems: An Application to Promotion Planning Problems

Paulo A. Pereira, Fernando A. C. C. Fontes, and Dalila B. M. M. Fontes

Abstract We propose a Hybrid Genetic Algorithm (HGA) for a combinatorial optimization problem, motivated by, and a simplification of, a TV Self-promotion Assignment Problem. Given the weekly self-promotion space (a set of TV breaks with known duration) and a set of products to promote, the problem consists of assigning the products to the breaks in the "best" possible way. The objective is to maximize contacts in the target audience for each product, whist satisfying all constraints. The HGA developed incorporates a greedy heuristic to initialize part of the population and uses a repair routine to guarantee feasibility of each member of the population. The HGA works on a simplified version of the problem that, nevertheless, maintains its essential features. The proposed simplified problem is a binary programming problem that has similarities with other known combinatorial optimization problems, such as the assignment problem or the multiple knapsack problem, but has some distinctive features that characterize it as a new problem. Although we are mainly interested in solving problems of large dimension (of about 200 breaks and 50 spots), the quality of the solution has been tested on smaller dimension problems for which we are able to find an exact global minimum using a branch-and-bound algorithm. For these smaller dimension problems we have obtained solutions, on average, within 1% of the optimal solution value.

keywords Genetic Algorithms, Combinatorial Optimization, TV Self-Promotion Assignment Problem.

Paulo A. Pereira
CMAT and Dept. of Mathematics and Applications, Universidade do Minho, 4800-058 Guimarães, Portugal, ppereira@math.uminho.pt

Fernando A. C. C. Fontes
ISR and Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal, faf@fe.up.pt

Dalila B. M. M. Fontes
LIAAD-INESC Porto L.A. and Faculdade de Economia, Universidade do Porto, 4200-464 Porto, Portugal, fontes@fep.up.pt

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Research Proceedings, DOI 10.1007/978-3-642-20009-0_48,

1 Introduction

In this article, we describe the development of an optimizer, based on genetic algorithms, which is integrated into a Decision Support System (DSS) to plan the best assignment of the weekly self-promotion space for a TV station. The DSS has been developed for SIC, a commercial TV station that frequently leads audience shares in Portugal. This article concentrates on describing the optimizer which is based on a hybrid Genetic algorithm (GA) combining a greedy heuristic and a constraint handling procedure based on linear programming relaxations.

Advertising is the main source of income in commercial TV's and advertisement revenues grow with audiences. Therefore TV stations typically reserve a promotion space (broadcasting time) for self-promotion (future programs, own merchandizing). It is also known that the total broadcasting time dedicated to advertisement, both commercial and self-promotion, is regulated and limited. Hence, the self-promotion time should be kept small, and optimizing its usage is of upmost importance. The problem is as follows: given the weekly self-promotion space (a set of breaks with known duration) and a set of products to promote, the aim is to assign the products to the breaks in the "best" possible way. For each product we have a given target audience, and for each break we can estimate how many viewers of each target are watching at one time. Moreover, each product has several assignment constraints to be satisfied. The objective is to maximize contacts in the target audience for each product, while satisfying all constraints.

The weekly problem faced by SIC, typically involves the broadcast of about 1350 spots, chosen from around 90 different ones, in roughly 200 time slots. The duration of each of these time slots, designated by breaks, ranges from 6 to 150 seconds. Given the problem complexity and dimensionality (see [4, 5] for a more detailed problem description), only heuristic approaches are expected to be able to obtain solutions in a reasonable amount of time. In the next section, we describe a GA that also incorporates a greedy heuristic to initialize part of the population with "good solutions". While the DSS considers the original problem with all the complexity features, the GA works on a simplified version of the problem that, nevertheless, maintains the essential features of the problem. We have noticed that the simplified problem is a binary programming problem that has similarities with other known combinatorial optimization problems, such as the multiple knapsack problem (see e.g. [6, 3]).

We propose a hybrid GA approach for the problem. First, we use a greedy heuristic procedure to construct good solutions. Each solution is a spot-break assignment binary matrix. Next, we handle the constraints by using a "repair" procedure to eliminate unfeasibilities. The population of solutions thus obtained is then evolved using the genetic operators reproduction, crossover, and mutation. The repair routine is used in each iteration of the GA to guarantee the feasibility of all solutions.

2 The Problem

A simplified version of the problem is stated as follows. Let $i = 1, \ldots, N_B$ be the set of breaks, and $j = 1, \ldots, N_S$ be the set of spots. Each break i and spot j are characterized by their durations B_i and d_j, respectively. Also, let c_{ij} denote the number of viewers within target that watch spot j when inserted in break i. We wish to find a spot-break assignment such that a maximum and minimum break usage (B_i and b_i respectively) for break i, as well as, a minimum number of viewers within target for spot j, C_j, are guaranteed. Amongst all possible assignments we are interested in the ones that maximize the number of viewers from the intended target. This problem can be formulated as given in the following binary programming model.

$$\text{Maximize} \sum_{i=1}^{N_B} \sum_{j=1}^{N_S} c_{ij} x_{ij} \tag{1}$$

$$\text{Subject to:} \sum_{j=1}^{N_S} d_j x_{ij} \leq B_i \quad \forall \ 1 \leq i \leq N_B, \tag{2}$$

$$\sum_{j=1}^{N_S} d_j x_{ij} \geq b_i \quad \forall \ 1 \leq i \leq N_B, \tag{3}$$

$$\sum_{i=1}^{N_B} c_{ij} x_{ij} \geq C_j \quad \forall \ 1 \leq j \leq N_S, \tag{4}$$

$$x_{ij} \in \{0,1\} \qquad \forall \ 1 \leq i \leq N_B \ \text{ e } \ \forall \ 1 \leq j \leq N_S, \tag{5}$$

where the decision variables x_{ij} assume the value 1 if spot j is assigned to break i and 0 otherwise.

We stress that this formulation, that is used in the GA, is a simplified version of the model that is used in the DSS, where additional problem features and constraints are dealt with.

3 The Hybrid Genetic Algorithm

We propose a GA that uses a greedy heuristic followed by a repair procedure to generate a population of feasible solutions. This population is then evolved through reproduction, crossover, and mutation. The reproduction and crossover operators determine which parents will have offspring, and how genetic material is exchanged between parents. The stochastic universal sampling (SUS) technique, introduced by [1], is used to select solutions for recombination. Such a selection is proportional to the fitness values and exhibits no bias and minimal spread. SUS uses a single random value to sample all of the solutions by choosing them at evenly spaced intervals. Crossover happens between a solution which is located in an even position with the solution located in the adjacent odd position.

The single point crossover is used. After selecting one crossover point, two children are obtained by using the binary string from the beginning of the chromosome to the crossover point from one parent, the rest being copied from the other parent. An elitist strategy is used, since some of the best individuals are copied from one generation to the next. This is important since, this way, we guarantee that the best solution is monotonically improved from one generation to the next. However, such a strategy tends to increase the convergence rate. Therefore, in order to avoid excessive convergence, mutation is also incorporated, through immigration, to allow for random introduction of new genetic material.

Greedy Heuristic:

We have developed a constructive greedy heuristic procedure that outputs spot-break assignment binary matrices with good coverage.

The procedure has 3 stages: In stage (i) some of the constraints of the DSS full model are converted into spot-constraints of type (4). Therefore, except for the break duration constraints which are dealt with differently, we only have spot-constraints. Although some constraints are easily converted, others have required complex procedures to do so. This is the case of the minimum coverage constraints. We have developed an estimator for the number of times each show would need to be advertised since the coverage is a nonlinear and unknown function of the broadcasted spots.

In stage (ii), using the minimum number of viewers C_j required for each spot j, we estimate (using a uniform assignment to the breaks) the maximum Mb_j and the minimum mb_j number of times that spot j may be used in order to satisfy (2) and (3), respectively .

Then, in stage (iii) we generate the binary matrix by resorting to an iterative greedy heuristic procedure. Spots are iteratively assigned to breaks, starting with the spot having the highest number of minimum times still to be used (initially mb_j). For such a spot, we select from the breaks that still have enough unused time, the one with the larger audience rate for the target of the selected spot. This is repeated until the coverage C_j has been provided for all spots, or the spot has already been used Mb_j times.

The Repair Routine:

In order to guarantee solution feasibility (unfeasible solutions are mainly due to the group of constraints that imposes not exceeding the breaks duration) we use a "repair" procedure, inspired in that of Chu and Beasley [2] for multiple knapsack problems. This procedure has two phases: in the first one we guarantee that the break time is not exceeded by removing spots, whenever needed, while in the second phase we try to satisfy spot number of viewers, or improve the solution, by adding spots to the breaks which have some unused time.

We start by selecting the most violated break, i.e. the break with the highest used time above its duration, and then iteratively remove from it spots in ascending order of u_{ij}, where

$$u_{ij} = \frac{c_{ij}}{w_i d_j}$$

and w_i is the dual value of the break constraint, obtained by solving a linear programming relaxation of the problem (1-4). Then, spots with a number of viewers smaller than C_j are further inserted into breaks with unused time, in ascending order of time usage benefit. This is obtained as the product of the spot duration by the dual variable of the break constraint ($w_i d_j$). If this is not possible, swaps with oversatisfied spots are attempted. At the end of this routine we have a feasible solution.

4 Computational Results

We have tested the proposed GA using 60 randomly generated problems of small dimension, consisting of 5 instances of 12 different sizes. Problem size is represented as (a, b), where a is the number of breaks and b the number of spots. The problem size (in the tests with small dimension problems) ranged from (6,3) to (8,8). The GA was used 10 times for each of these problems, therefore in the following results each reported figure corresponds to the average value obtained from the 10 runs of the 5 problem instances. In Figure 1 we report on the solution quality and computational time requirements.

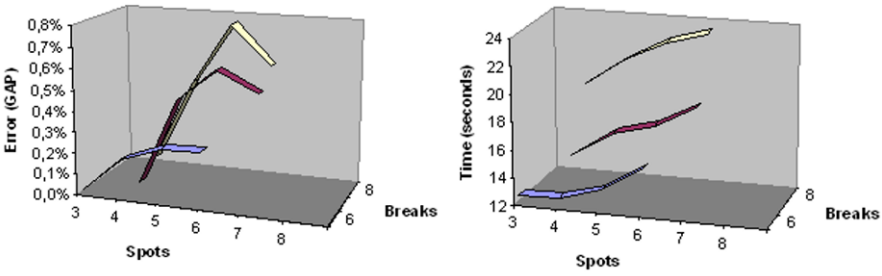


Fig. 1 Computational performance for randomly generated problems.

The time requirements are reported in seconds and the solution quality is the optimality gap obtained as $\frac{GA_{sol}-OPT}{OPT}$, where OPT is the optimal solution value obtained by solving the problems using a branch-and-bound algorithm.

From the analysis of Fig. 1, we conclude that, as expected, both the computational time and the gap increase with problem size. However, for some problems with the same number of breaks, there is a decrease in the gap value with the increase of the number of spots. This can be explained by a decrease in the total number of admissible solutions when the number of spots approaches the number of breaks. All solutions are, on average, within 1% gap.

We have also solved 3 problem instances with 200 breaks and 50 spots, which is the size of the TV station real problems. However, for such problems we can only report on computational time requirements since the branch-and-bound algorithm

only solves problems with up to 8 breaks. Again the results reported refer to the average of the 10 runs performed using the GA.

Size (200,50)	Best Sol. Value	Best Sol. Generation	Best Sol. Time
Problem 1	107962.30	353.35	20692.30
Problem 2	112203.30	398.35	20791.30
Problem 3	107803.00	352.70	20782.80
Average		368.13	20755.50
Std. Dev.		21.40	44.80

Table 1 Computational performance for real sized problems.

In Table 1, we report for each of the problems the value of the objective function (the maximum total number of contacts), in the first column. In the second column, we report the number of generations after which the solution converges. In the last columns, we report the computational time in seconds. We can see that the computational time is, on average, about 5h45m. This value is adequate for a planning that is carried out weekly and compares very favorably with previous operator generated solutions.

Acknowledgements We acknowledge the support of SIC and project FCT PTDC/EGE/GES/099741/08.

References

1. J.E. Baker. Reducing bias and inefficiency in the selection algorithms. In *Proceedings of the Second International Conference on Genetic Algorithms*, pages 14–21. Lawrence Erlbaum Associates Inc., USA, 1987.
2. J.E. Chu, P.C. e Beasley. A genetic algorithm for the multidimensional knapsack problem. *Journal of Heuristics*, 4: 63–86, 1998.
3. E.Y.H. Lin. A bibliographical survey on some well-known non-standard knapsack problems. *INFOR*, 36(4): 274–317, 1998.
4. P.A. Pereira, F.A.C.C. Fontes, and D.B.M.M. Fontes. A Decision Support System for Planning Promotion Time Slots. In *Operations Research Proceedings 2007: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR)*, page 147. Springer, 2008.
5. Paulo A. Pereira. *Um Sistema de Apoio à Decisáo Baseado em Optimizaçáo para o Planeamento de Auto-promoçáo de uma Estaçáo de Televisáo*. PhD thesis, Departamento de Matemática para a Ciência e Tecnologia, Universidade do Minho, Guimaráes, Portugal, 2009.
6. J. Tavares, FB Pereira, and E. Costa. Multidimensional knapsack problem: A fitness landscape analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 38(3): 604–616, 2008.

Route Planning for Robot Systems*

Martin Skutella and Wolfgang Welz

Abstract In welding cells a certain number of robots perform spot welding tasks on a workpiece. The tours of the welding robots are planned in such a way that all weld points on the component are visited and processed within the cycle time of the production line. During this operation, the robot arms must not collide with each other and safety clearances have to be kept. On the basis of these specifications, we concentrate on the Discrete Optimization aspects of the stated problem. This leads to a Vehicle Routing based problem with additional scheduling and timing aspects induced by the necessary collision avoidance. This problem can then be solved as an Integer Linear Program by Column Generation techniques. In this context, we adapt the Resource Constrained Shortest Path Problem, so that it can be used to solve the pricing problem with collision avoidance. Using this approach, we can solve generated test instances based on real world welding cells of reasonable size.

1 Introduction and Problem Formulation

Welding cells are essential elements in many production systems of the automobile industry. In these cells, several robots perform welding tasks on the same component. For instance, four robots have to make 50 weld points on a car door. The movements of the welding robots need to be planned in such a way that all weld points on the component are visited and processed within the cycle time of the production line. The robot arms possess different properties: They are equipped with different welding tongs and can only process specific weld points. During this operation the robot arms must not collide with each other or the component. Given the CAD data of the workpiece, the task is to find a feasible sequence of nodes as well

Institute of Mathematics MA 5-2, TU Berlin, Str. des 17. Juni 136, 10623 Berlin, Germany, e-mail: {skutella,welz}@math.tu-berlin.de

* Supported by DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_49,

as the path planning for all the robots. A somehow related laser welding problem where robots have to share a limited number of laser sources is considered in [2, 6].

To be able to solve this problem with Discrete Optimization methods, we consider a simplification: We first assume that the time a robot needs to travel from one weld point to another is given and, in particular, does not depend on the orientation of the robot arm or other factors. Further, we also consider a slightly simplified notion of collision detection: For collision avoidance only the paths the robots use between two points are taken into account as a whole and not the actual position or orientation of the arms during this motion process.

For a formal definition of the resulting problem, the following needs to be given:

- A set of nodes V consisting of all the weld points and a starting node $s_k \in V$ for each robot $k = 1, \ldots, K$.
- A set of arcs A_k for each robot. An arc $(u,v) \in A_k$ means that robot k can move from node v to node u.
- For each arc $a \in A_k$, we have the traversal time $t_a^k > 0$ for this arc and the cost c_a^k.
- An upper bound L on the traversal time of a tour. This time represents the cycle time of the current welding cell that must not be exceeded.
- A collision function $col : \{1, \ldots, K\} \times V \times V \times \{1, \ldots, K\} \times V \times V \to \{0,1\}$. We have $col(k,u,v,k',u',v') = 0$, if there are no conflicts between robot k going from u to v and robot k' going from u' to v' at the same time. Thus, $col(k,u,v,k',u',v') = 1$ if the corresponding robot motions are not compatible.

To check whether two tours collide, we add a certain notion of time to the tours: A *scheduled tour* is a sequence of arcs together with start and end times for each of these arcs. For a tour arc a we denote its start time with $start_a$ and its end time with end_a. All of these times are at least zero and discretized (i. e., in $\mathbb{N}_{\geq 0}$). For all arcs $a \in A_k$ it holds that $end_a - start_a = t_a^k$. If arc a' is the exact successor of arc a in the same tour, we always have $end_a \leq start_{a'}$. If end_a is strictly smaller than $start_{a'}$, we say that we have a *waiting time* of $start_{a'} - end_a$ at the corresponding node.

We call the two scheduled tours T^k of robot k and $T^{k'}$ of robot k' *collision free* if and only if $col(k, a_{source}, a_{target}, k', a'_{source}, a'_{target}) = 0$ holds for all pairs of arcs a, a', where a is an arc of tour T^k, a' is an arc of $T^{k'}$, and a, a' are in use at the same time (i. e., $start_a \leq start_{a'} < end_a + w_a$ or $start_{a'} \leq start_a < end_{a'} + w_{a'}$, where w denotes the waiting time in the target node of the corresponding arc). Also the waiting times are crucial for the compatibility of tours, because even if a robot waits in a node, other robots can collide with it.

The *Welding Cell Problem* (WCP) is the problem of finding a scheduled tour for each robot with minimal total cost, so that every vertex is visited. The individual scheduled tours must not exceed the given cycle time and must be pairwise collision free. When the collision constraints are dropped, we obtain a *Vehicle Routing Problem* (VRP). Classical exact approaches for solving large VRPs are based on Integer Programming and Column Generation. An introduction to this technique can be found, e. g., in [1].

2 Column Generation Approach

The WCP can be modeled as an extended *Set Partitioning Problem* and formulated as a Binary Linear Program with additional constraints for the collision avoidance:

$$\begin{aligned} \min \quad & \sum_{\theta \in \Omega} c_\theta x_\theta \\ \text{s.t.} \quad & \sum_{\theta \in \Omega} \delta_{v\theta} x_\theta = 1 && \forall v \in V && (1) \\ & x \text{ is collision free} && && (2) \\ & x_\theta \in \{0,1\} && \forall \theta \in \Omega && (3) \end{aligned}$$

The set Ω is defined as $\Omega := \bigcup_{k \in \{1,\ldots,K\}} \Omega_k$ with Ω_k being the set of all feasible tours for robot k. The coefficient $\delta_{v\theta}$ is 1 if the node v is visited by tour θ and 0 otherwise.

The set partitioning constraints (1) guarantee that every node is visited exactly once. All other problem constraints are implicitly given by the definition of Ω_k: The set Ω_k contains only those tours that visit the starting node and that have a traversal time within the given time limit.

By relaxing the constraints (2) and (3) to just $x_\theta \geq 0$ for all $\theta \in \Omega$, we get a pure Linear Program with exponentially many variables, where standard column generation techniques can be applied.

The reduced cost for variable x_θ in this formulation, with π_v being the dual variables corresponding to the constraints (1), can be rewritten as an arc-based expression:

$$c_\theta - \sum_{v \in V} (\delta_{v\theta} \pi_v) = \sum_{a \in A_k} x^k_{\theta a} c^k_a - \sum_{a=(u,v) \in A_k} x^k_{\theta a} \pi_v \,,$$

where $x^k_{\theta a}$ is 1 if arc a is in the tour θ of robot k, and 0 otherwise.

Finding the variable x_{θ^*} with minimal reduced cost, the so called pricing problem, corresponds to the combinatorial optimization problem of finding the shortest tour θ^* with respect to the arc costs modified by π. Tour θ^* must be a valid tour for robot k, which means that it must visit the starting node s_k and that no vertex is visited more than once, i. e., the tour is elementary. Additionally, the traversal time of the tour must be bounded by L.

Since the dual variables π_v correspond to equality constraints in the primal problem, they can become negative and it is therefore possible that negative cost cycles exist.

Enforcing the integrality constraints in Integer Linear Programs is usually handled with a branch-and-bound approach. Branching can also be applied to solve the given formulation for the WCP: Since the branching constraints must also be considered in the pricing, it is the more direct way to branch on arcs because our pricing is arc-based.

We choose a tour with fractional value and on this tour we select an arc whose value, i. e., the sum of the values of all tours using this arc, is less than one and

therefore fractional. Since the starting nodes s_k cannot be visited by any other robot than k and due to the equality constraints for the nodes, such an arc always exists. We then branch on this selected arc a. For the first branch, we force arc $a = (u, v)$ to be in the tour of one robot. This can be done by simply removing all other arcs leaving the node u and all arcs entering v. The banned arcs stay removed for the entire sub-tree in the branch-and-bound tree, but eventually additional arcs are also removed for other child nodes. For the second subproblem, we remove arc a in all graphs of the corresponding sub-tree. If the values of all arcs are either 0 or 1, this induces binary values for all tour variables and therefore the constraints (3). All these branching decisions can be modeled by changes in the problem graph of the corresponding pricing while the problem itself stays unchanged as no explicit constraints are added.

Also the constraint (2) can be handled with branch-and-bound. Assume we are using two incompatible arcs a_1 and a_2 simultaneously in the time interval $[t_a, t_e]$. We then select one time step $t_b \in [t_a, t_e]$ and branch on this time step. For time step t_b, there are three compatible possibilities for the arcs a_1 and a_2:

(i) a_1 is used and a_2 is not
(ii) a_2 is used and a_1 is not
(iii) neither a_1 nor a_2 are used

So we create three subproblems where in each one of them one of these possibilities is ensured by forcing or forbidding the corresponding arcs in time step t_b. We say an arc a is *forced in time step* t_b if the arc must be used during t_b. It does not matter whether a is used in other time steps or not. This branching scheme is valid because every collision free integer solution always fits into one of the branches and, therefore, no solution is lost. Moreover the branching scheme will find the optimal collision free integer solution in finitely many steps, since there are only finitely many possible time steps to branch on and only finitely many arcs.

To prevent the generation of duplicate tours and to ensure the optimality of the pricing problem, these constraints must also be imposed for the pricing. This then leads to the problem of finding the shortest elementary tour with additional multiple time windows on the arcs. However, since the traversal times of the arcs are known in advanced, these time windows can easily be transferred into time windows on the nodes.

3 Combinatorial Approach for the Pricing Problem

The resulting pricing problem can be interpreted as an instance of the so called *Resource Constrained Shortest Path Problem* (RCSPP). The RCSPP seeks a shortest (not necessarily elementary) path in a directed graph with arbitrary arc costs from a starting node to a target node under certain "resource constraints". A survey on the problem can be found, e. g., in [3].

The standard solution approach for RCSPP is a labeling algorithm based on dynamic programming techniques. The general idea is very similar to the Shortest Path Problem without resource constraints: All possible partial paths are extended arc by arc until a shortest path is found. However, in contrast to the Shortest Path Problem without resource constraints, two paths of the RCSPP can be incomparable. This labeling approach uses the concepts of resources and resource extension functions. A resource r is a one-dimensional quantity that is consumed along a path and can be evaluated at every node of the path. The consumption of resource r along arc a is described by a non-decreasing function f_a^r, the so called *Resource Extension Function*. The resource constraints itself are expressed by resource windows at each node, these intervals contain the allowed values for the resource at this node.

Labels are used to store information about partial paths. A label consists of its current vector of resource values and its predecessor arc. The labels belong to a node and the resource extension functions are used when they are extended along an arc.

An extension of a label along an arc is feasible, if the resource values in the resulting label do not exceed their resource windows. If the value of a resource is less than the lower endpoint of the corresponding resource window, it must be always possible to render this label feasible by consuming an extra amount of this resource without moving along an arc. If the corresponding value belongs to the resource *time*, this can be interpreted as waiting in the current node.

To keep the number of labels as low as possible, it is important to identify and then discard unnecessary labels with the so called *dominance rules*: Given two different feasible labels ℓ and ℓ' in the same node, ℓ dominates ℓ' if $R(\ell) \leq R(\ell')$, where R is the vector of the resource values. Because of the requirements of the RCSPP, no optimal solution is hereby lost.

In addition to that, the approach can be extended to only generate elementary tours: For each node $v \in V$ an additional resource r_v is added. This resource must either be 0 or 1 and counts the number of times the node v was visited in the current partial path. This is further improved by using a slightly more complicate dominance rule that already eliminates paths that contain 2-cycles as described in [4, 5].

All of these techniques can be applied to solve our pricing problem using the resources *cost* representing the reduced cost of the tour and *time* for the start time of the next arc in the current node. However, the pricing problem is no RCSPP of the given definition, as in our case waiting in a node is not always allowed, since even a waiting robot can cause collisions. Under these assumptions, it no longer holds that needed labels will not be discarded by the given dominance rule (see Fig. 1). Hence, the dominance rules have to be adapted: Waiting can only affect the resource *time*. If time is not taken into account during the dominance rules, this problem is resolved. However, many unnecessary labels are still kept. To improve this number, we calculate a critical time value t_v^{crit} for each node v, so that a label with a time value greater than t_v^{crit} cannot be extended to paths that have to wait in any of their nodes after v. If the value is smaller, waiting times can occur. The point in time t_v^{crit} can be calculated in advance based on the resource windows and these values do not change during the labeling algorithm. This can then be used for an improved dominance rule: ℓ dominates ℓ' if $R(\ell) \leq R(\ell')$ and either $R^{time}(\ell) = R^{time}(\ell')$ or $R^{time}(\ell) > t_v^{crit}$.

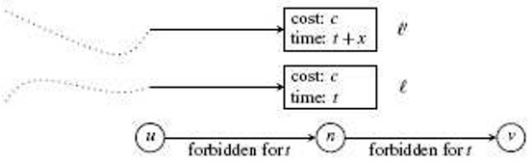


Fig. 1 Label ℓ dominates ℓ', both using the arc (u,n). Label ℓ is feasible in n but cannot be extended to v as it cannot wait in n. Label ℓ' can be extended if x is big enough. Hence, ℓ' must not be discarded.

4 Results and Conclusion

The presented approach was tested on randomly generated test instances based on real world welding cell problems containing four robots processing 10 to 40 weld points. Using a 2.6 GHz AMD Opteron Processor with 1024 KB cache and a memory limit of 2 GB together with a time limit of 12 hours, instances with up to 31 weld points could be solved. Of the 21 test instances with reasonable sizes of 22–31 points, we could find a feasible solution for 19 instance. Out of these, the algorithm could show optimality for 15 instances.

The verification of these results on real-world welding data is currently being done as part of project "Automatic reconfiguration of robotic welding cells" within the DFG Research Center MATHEON in Berlin. Within this project, the nonlinear optimization problems of robot motion planing and collision detection shall be integrated in the presented approach.

Acknowledgements The authors are much indebted to Ralf Borndörfer, Martin Grötschel, René Henrion, Dietmar Hömberg, Jörg Rambau and Cornelius Schwarz for helpful discussions on the topic of this paper.

References

1. J. Desrosiers and M.E. Lübbecke. A primer in column generation. In G. Desaulniers, J. Desrosiers, and M. M. Solomon, editors, *Column Generation*, chapter 1, pages 1–32. Springer, 2005.
2. M. Grötschel, H. Hinrichs, K. Schröer, and A. Tuchscherer. A mixed integer programming model for a laser welding problem in car body manufacturing. Technical report 06–21, Zuse Institute Berlin, 2006.
3. S. Irnich and G. Desaulniers. Shortest path problems with resource constraints. In G. Desaulniers, J. Desrosiers, and M. M. Solomon, editors, *Column Generation*, chapter 2, pages 33–65. Springer, 2005.
4. N. Kohl. *Exact methods for time constrained routing and related scheduling problems*. PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 1995.
5. J. Larsen. *Parallelization of the Vehicle Routing Problem with Time Windows*. PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 1999.
6. Jörg Rambau and Cornelius Schwarz. On the benefits of using NP-hard problems in branch & bound. In *Proceedings of the International Conference Operations Research 2008*, pages 463–468. Springer, 2009.

On Universal Shortest Paths

Lara Turner and Horst W. Hamacher

Abstract The universal combinatorial optimization problem (Univ-COP) generalizes classical and new objective functions for combinatorial problems given by a ground set, a set of feasible solutions and costs assigned to the elements in the ground set. The corresponding universal objective function is of the sum type and associates additional multiplicative weights with the ordered cost coefficients of a feasible solution such that sum, bottleneck or balanced objectives can, for instance, be modeled. For the special case of shortest paths, we give two alternative definitions for the corresponding universal shortest path problem denoted Univ-SPP, one based on a sequence of cardinality constrained subproblems, the other using an auxiliary construction to establish uniform length for all paths from s to t. We show that the second version can be solved as classical sum shortest path problem on graphs with specific assumptions on edge costs and path lengths. In general, however, the problem is NP-hard. Integer programming formulations are proposed.

1 Introduction

The shortest path problem (SPP) is one of the best known models in mathematical optimization with a huge number of publications. One of the reasons for this is its potential in tackling real-world problems, another its usage as subproblem in larger optimization problems.

In general, one assumes that SPP is the sum-SPP, where paths are compared with respect to the sum $\sum_{e\in P} c(e)$ of the edge costs in a given directed graph. Another

Lara Turner
Technical University of Kaiserslautern, Department of Mathematics, P.O. Box 3049, 67653 Kaiserslautern, Germany, e-mail: turner@mathematik.uni-kl.de

Horst W. Hamacher
Technical University of Kaiserslautern, Department of Mathematics, P.O. Box 3049, 67653 Kaiserslautern, Germany, e-mail: hamacher@mathematik.uni-kl.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_50,

well-known SPP is the bottleneck SPP in which the maximum edge cost $\max_{e\in P} c(e)$ along P is minimized. Other versions of SPPs include balanced SPP (minimizes the difference between largest and smallest edge cost), k-max SPP or k-sum SPP (Garfinkel et al [2]) where the latter two determine a path in which the k^{th}-largest edge cost or the sum of the k largest edge costs is as small as possible. More exotic SPPs are the (k,l)-balanced SPP (with minimal difference between k^{th}-largest and l^{th}-smallest edge cost), the $(k+l)$-max SPP (minimizing the sum of k^{th}-largest and l^{th}-largest edge cost) and the (k,l)-trimmed-mean SPP in which the k largest and l smallest edge costs are ignored and the costs of the remaining edges are added. These problems are addressed in Turner [5].

In the past, several attempts have been made to develop unifying theories with the goal of finding common properties and solution algorithms for a large class of SPPs. Most notably, this was achieved by the class of algebraic path problems (see e.g. Zimmermann [8]), where the sum or max operation in the objective function is replaced by a more general bi-linear operator and the real numbers are replaced by the elements of a semigroup or semiring. In this paper, however, we choose a different approach where the objective function is of the sum type and multiplicative weights are used to model various instances of shortest path problems. It is based on the ideas of Yager [7] for ordered weighted averaging aggregation operators in multicriteria decisionmaking and those of Nickel and Puerto [3] for location problems with ordered median function.

2 Universal Combinatorial Optimization Problem

A combinatorial optimization problem (COP) is given by a finite set E (ground set) of cardinality m, a family of subsets $F \subseteq 2^E$ (feasible solutions) and a cost function $c : E \to \mathbb{R}$ assigning costs $c(e) \in \mathbb{R}$ to the elements in E. The most popular objective functions of combinatorial problems are the sum and bottleneck objective which are special cases of the more general problem introduced next.

Definition 1 *Consider a COP in which all feasible solutions $S \in F$ have the same cardinality $|S| \leq m$. For any $S \in F$ we denote with $c_{(i)}(S), i = 1,\ldots,|S|$, its i^{th}-largest cost coefficient, i.e. $c_{(1)}(S) \geq \ldots \geq c_{(|S|)}(S)$, and with $c_{\geq}(S) := (c_{(1)}(S),\ldots,c_{(|S|)}(S))$ its **vector of sorted costs**. Given weights $\lambda_i \in \mathbb{R}, i = 1,\ldots,|S|$, the **universal combinatorial optimization problem, Univ-COP,** is*

$$\min_{S\in F} f_\lambda(S) := \sum_{i=1}^{|S|} \lambda_i c_{(i)}(S). \tag{1}$$

To call such problems universal makes truly sense, since, by choosing appropriate weights λ_i, classical and new objective functions can be represented by (1). Objective functions which can be modeled are, for instance, the standard sum or bottleneck objective for which we set $\lambda_i = 1$ for all $i = 1,\ldots,|S|$ or $\lambda_1 = 1$ and $\lambda_i = 0$ otherwise. The balanced objective function $\max_{e\in S} c(e) - \min_{e\in S} c(e)$ is formulated

by setting $\lambda_1 = 1, \lambda_{|S|} = -1$ and $\lambda_i = 0$ else. A further generalization of the balanced objective is given in Turner et al [6]. k-max and k-sum objective functions are obtained for $\lambda_k = 1$ or $\lambda_1 = \ldots = \lambda_k = 1$ and the remaining $\lambda_i = 0$.

An example of such a Univ-COP is the universal spanning tree problem (see also Fernández et al [1]) for which the greedy algorithm can be shown to find an optimal solution if $\lambda_i \geq 0$ for all $i = 1, \ldots, |S|$. The same is true for the universal matroid base problem.

3 Universal Shortest Path Problem

The classic single-source single-sink shortest path problem is a special type of a COP where the ground set is given by the edges of a directed graph $G = (V, E)$, the feasible solutions are the paths from source s to sink t and costs $c(e) \in \mathbb{R}$ are assigned to the edges $e \in E$. The set of all (s,t)-paths in G is denoted by $\mathcal{P}_{st}$ and, by definition, a path is a sequence of nodes and edges such that neither nodes nor edges are repeated. Thus, a path $P \in \mathcal{P}_{st}$ has at most $n-1 = |V|-1$ edges and this number of edges is called the length $l(P)$ of path P. We may assume that there are no incoming edges in s and no outgoing edges from t in graph G.

The crucial assumption in Definition 1, that the cardinality of the feasible solutions is fixed, is, in general, not satisfied when considering shortest paths. The fact that Definition 1 cannot be used to model a universal shortest path problem (Univ-SPP) is e.g. illustrated by the balanced SPP. In order to make its objective a special case of (1), one would have to set $\lambda_1 = 1$ in order to reproduce the first part $\max_{e \in P} c(e)$ of the objective function. To reproduce the second part $-\min_{e \in P} c(e)$ two different weight vectors with components $\lambda_l = -1, \lambda_{l'} = 0$ and $\lambda_l = 0, \lambda_{l'} = -1$ would be required to compare two paths with different number of edges l and l'.

3.1 Univ-SPP – Sequential Definition

The first definition of Univ-SPP which we propose in this section consists of $n-1$ subproblems with fixed cardinality for all paths from s to t.

Definition 2 *For any path $P \in \mathcal{P}_{st}$ of length $l(P) = l$ with $l \in \{1, \ldots, n-1\}$, the **sorted cost vector** of path P sorts the costs along P in non-increasing order, i.e. $c^l_{\geq}(P) := (c^l_{(1)}(P), \ldots, c^l_{(l)}(P))$ where $c^l_{(i)}(P), i = 1, \ldots, l$, is the i^{th}-largest edge cost in path P.*

Definition 3 *Given weight vectors $\lambda^l \in \mathbb{R}^l$ for all $l \in \{1, \ldots, n-1\}$, the **sequential universal shortest path problem, Univ-SPP($\mathbf{1}, \ldots, \mathbf{n-1}$),** is defined as*

$$\min_{l \in \{1,\ldots,n-1\}} f_{\lambda^l}(P^*_l).$$

P_l^* is a path from s to t with length l which is optimal to the ***universal shortest path problem with cardinality*** **l**, ***Univ-SPP(*****l****)***,

$$\min_{P\in\mathcal{P}_{st}:l(P)=l} f_{\lambda^l}(P) := \sum_{i=1}^{l} \lambda_i^l c_{(i)}^l(P)$$

where the set of feasible solutions is restricted to those paths $P \in \mathcal{P}_{st}$ with length l.

Choosing weight vectors λ^l for any path length $l \in \{1,\dots,n-1\}$, sum- and bottleneck SPP are modeled in an obvious way. Setting $\lambda_1^l = 1, \lambda_l^l = -1$ and $\lambda_i^l = 0$ else for $l \neq 1$ and $\lambda_1^1 = 0$ or $\lambda_1^1 = \infty$ for paths consisting of only one edge (depending on whether they are assumed to be feasible or not) yields the balanced SPP. Similarly for objective functions like k-max or k-sum that require a minimum path length k to be well-defined we will define $\lambda_i^l = 0$ or $\lambda_i^l = \infty$ for all $l < k$ and $i = 1,\dots,l$.

Since Univ-SPP$(n-1)$ includes the Hamiltonian path problem, it follows:

Theorem 1 *Univ-SPP$(1,\dots,n-1)$ is (strongly) NP-hard.*

3.2 Univ-SPP with Cardinality $|E|$

For the case of non-negative costs $c(e) \geq 0$ we suggest a second definition for Univ-SPP. The main idea is to enforce for all paths from s to t a length of $m = |E|$ where m is the number of edges in graph G. This is achieved by extending each (s,t)-path in the original graph by artificial edges of cost 0.

Definition 4 *The* ***extended sorted cost vector*** *of a path $P \in \mathcal{P}_{st}$ is given as*

$$c_{\geq}(P) := (c_{(1)}(P),\dots,c_{(l(P))}(P),\underbrace{0,\dots,0}_{m-l(P)})$$

where $c_{(i)}(P), i = 1,\dots,l(P)$, is the i^{th}-largest edge cost in P, i.e. $c_{(1)}(P) \geq \dots \geq c_{(l(P))}(P) \geq 0$, and $c_{(i)}(P) := 0$ for all $i = l(P)+1,\dots,m$.

Definition 5 *For a weight vector $\lambda \in \mathbb{R}^m$, the* ***universal shortest path problem with cardinality*** **$|$E$|$ = m**, ***Univ-SPP(*****m****)***, *is*

$$\min_{P\in\mathcal{P}_{st}} f_\lambda(P) := \sum_{i=1}^{m} \lambda_i c_{(i)}(P). \tag{2}$$

Using that the edge costs $c(e), e \in E$, are non-negative and all artificial costs are equal to 0, it can be verified that sum-SPP and bottleneck SPP can be modeled correctly. The k-sum SPP is a special case of (2) by setting $\lambda_1 = \dots = \lambda_k = 1, \lambda_{k+1} = \dots = \lambda_m = 0$. This definition assigns paths with length $l(P) < k$ the total sum of their edge costs as objective value. For k-max SPP, we choose $\lambda_k = 1$ and $\lambda_i = 0$ for all $i \neq k$ and obtain an optimal objective value of 0 if a path of length less than k exists.

In contrast to sum-SPP which can be solved in polynomial time if all edge costs are non-negative, Univ-SPP(m) turns out to be NP-hard. This can be shown by reduction from the longest path problem.

Theorem 2 *Univ-SPP(m) is (strongly) NP-hard.*

Another difference compared with sum-SPP is that for Univ-SPP(m) the optimality principle of dynamic programming which guarantees, in case of sum-SPP, the correctness of algorithms as Dijkstra or Bellman-Ford is no longer valid. This motivates the search for more sophisticated solution methods and the investigation of special cases. We start with the simple observation of Lemma 1.

Lemma 1 *Let $\lambda_i \geq 0$ for all $i = 1,\ldots,m$ and let $P,P' \in \mathcal{P}_{st}$ with $c_{(i)}(P) \leq c_{(i)}(P')$ for all $i = 1,\ldots,m$. Then $f_\lambda(P) \leq f_\lambda(P')$.*

An example showing that there may be exponentially many incomparable extended sorted cost vectors is given in [5]. In the following, however, a shortest (s,t)-path P^* with respect to (modified) sum objective always has the property that $c_{\geq}(P^*) \leq c_{\geq}(P)$. A universal shortest path can thus be found in polynomial time.

Theorem 3 *Given $\kappa,\kappa' \in \mathbb{R}_+$ and non-negative weights $\lambda_i \geq 0$ for all $i = 1,\ldots,m$, any optimal solution of sum-SPP is a universal shortest path if*

1. *the costs are uniform, i.e. $c(e) = \kappa$ for all $e \in E$, or*
2. *the costs are binary, i.e. $c(e) \in \{0,\kappa\}$ for all $e \in E$, or*
3. *$c(e) \in \{\kappa,\kappa'\}$ for all $e \in E$ and all paths $P \in \mathcal{P}_{st}$ have the same length.*

In layered graphs where the node set can be partitioned into layers $V_0 = \{s\},\ldots,V_l = \{t\}$ and all edges run between consecutive layers, any path from s to t has length l and Univ-SPP(m) can be tackled as follows:

Theorem 4 *For a layered graph with monotone (s,t)-costs, i.e. $c(e_1) \geq \ldots \geq c(e_l)$ or $c(e_1) \leq \ldots \leq c(e_l)$ for all (s,t)-paths $P = (s = i_0,e_1,i_1,\ldots,i_{l-1},e_l,i_l = t)$, and $\lambda \in \mathbb{R}^m$, Univ-SPP(m) can be solved as sum-SPP with respect to modified costs $\lambda_k c(e)$ or $\lambda_{l-k+1}c(e)$ for all $e = (i,j) \in E$ with $i \in V_{k-1}$, $j \in V_k$ and $k \in \{1,\ldots,l\}$.*

To generalize this result to digraphs of arbitrary structure, we define the expanded graph G' which is a layered graph obtained from G by including $m+1$ copies $i(0),\ldots,i(m)$ of each node $i \in V$, m copies $(i(0),j(1)),\ldots,(i(m-1),j(m))$ of each edge $(i,j) \in E$ and m edges of type $(t(k),t(k+1))$ for $k = 0,\ldots,m-1$. For the costs, we set $c'(i(k),j(k+1)) = c_{ij}$ and $c'(t(k),t(k+1)) = 0$ for all $k = 0,\ldots,m-1$. Identifying nodes $s(0)$ and $t(m)$ with the source and sink of G', there is an 1:1-correspondence between (s,t)-paths in G and $(s(0),t(m))$-paths in G' if the digraph G is acyclic. Together with the following lemma this proves Theorem 5.

Lemma 2 *Let W be an (s,t)-walk in G (i.e. a non-simple path in which node and edge repetition is allowed) and let P be the (s,t)-path obtained from W by eliminating all cycles. If $\lambda_i \geq 0$ for all $i = 1,\ldots,m$, we have that $f_\lambda(P) \leq f_\lambda(W)$.*

Theorem 5 *Univ-SPP(m) in G and G' are equivalent if G is acyclic or $\lambda_i \geq 0$ for all $i = 1,\ldots,m$. If the costs are, in addition, non-increasing along (s,t)-paths and (s,t)-walks, Univ-SPP(m) is solvable as sum-SPP using modified costs as in Theorem 4.*

4 Integer Programming Formulations

Unlike sum-SPP, Univ-SPP(m) cannot be solved as linear program since the additional sorting problem makes the objective function non-linear. The variables x_{ij} are the path variables and the sorting of the edge costs is organized by variables $s_{k,ij}$:

$$\begin{aligned}
\min \quad & \sum_{k=1}^{m} \lambda_k \sum_{(i,j)\in E} s_{k,ij} c_{ij} x_{ij} \\
\text{s.t.} \quad & \sum_{k=1}^{m} s_{k,ij} = 1 && \forall (i,j) \in E \\
& \sum_{(i,j)\in E} s_{k,ij} = 1 && \forall k = 1,\dots,m \\
& \sum_{(i,j)\in E} s_{k,ij} c_{ij} x_{ij} \geq \sum_{(i,j)\in E} s_{k+1,ij} c_{ij} x_{ij} && \forall k = 1,\dots,m-1 \\
& \sum_{j\in\delta^+(i)} x_{ij} - \sum_{j\in\delta^-(i)} x_{ji} = \begin{cases} 1 & \text{if } i = s \\ -1 & \text{if } i = t \\ 0 & \text{if } i \neq s,t \end{cases} \\
& \sum_{(i,j)\in E(S)} x_{ij} \leq |S| - 1 && \forall S \subseteq V \\
& s_{k,ij}, x_{ij} \in \{0,1\} && \forall k = 1,\dots,m, (i,j) \in E.
\end{aligned}$$

A mixed integer linear programming formulation for Univ-SPP(m) is based on Ogryczak and Tamir [4]. It is valid for non-increasing and non-negative weights only. Both programs can be easily modified to describe the subproblems Univ-SPP(l) of the sequential definition Univ-SPP($1,\dots,n-1$). For the polyhedral analysis of the IP-formulations and their linearizations, we refer to [5].

References

1. E. Fernández, J. Puerto, and A.M. Rodríguez-Chía. On discrete optimization with ordering. 2008. Unpublished.
2. R. Garfinkel, E. Fernández, and T.J. Lowe. The k-centrum shortest path problem. *Top*, 14(2): 279–292, 2006.
3. S. Nickel and J. Puerto. *Location theory: A unified approach.* Springer, 2005.
4. W. Ogryczak and A. Tamir. Minimizing the sum of the k largest functions in linear time. *Information Processing Letters*, 85(3): 117–122, 2003.
5. L. Turner. PhD thesis, Technical University of Kaiserslautern, 2011. To appear.
6. L. Turner, A.P. Punnen, Y.P. Aneja, and H.W. Hamacher. On generalized balanced optimization problems. 2010. To appear in Mathematical Methods of Operations Research.
7. R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1): 183–190, 1988.
8. U. Zimmermann. *Linear and combinatorial optimization in ordered algebraic structures*, volume 10 of *Annals of Discrete Mathematics*. North-Holland, 1981.

P-Median Problems with an Additional Constraint on the Assignment Variables

Paula Camelia Trandafir and Jesús Sáez Aguado

Abstract We consider P-median problems with an additional constraint on the assignment variables. In this paper we develop two kinds of heuristic algorithms. Firstly, by designing both construction and improvement heuristics, which extend the Teitz-Bart and Whitaker heuristics to the case of an additional constraint, we develop a multistart GRASP heuristic. Secondly, by employing Lagrangian relaxation methods, we propose a very efficient heuristic, based on solving heuristically several unrestricted P-median problems. We present the computational results of comparing the different approximations using a set of randomly generated test instances.

1 Introduction

The P-median problem (PMP) and the maximum covering location problem (MCLP) are two of the most important location problems, both with a lot of applications (see ([1]). Meanwhile the goal of PMP is to minimize the total distance, the aim of MCLP is to maximize the demand covered at a distance lower or equal than some pre-established covering distance d_C. In this paper we consider an extension of PMP obtained adding a covering constraint (CONPMP). The restriction establishes that the total covered demand at a distance greater than d_C should not exceed a previously chosen value ε. While the solution of PMP is associated to the maximum efficiency when designing a service, a solution of the CONPMP looks for maximum efficiency subject to a given minimum required quality.

We shall develop different heuristic procedures for the CONPMP. The first group of algorithms are modified versions of the local search and exchange algorithms used in the PMP ([7], [8]). Finally the result is an improvement heuristic as well as a multistart GRASP heuristic for the CONPMP.

Paula Camelia Trandafir · Jesús Sáez Aguado
University of Valladolid, Department of Statistics and Operations Research, Spain, e-mail: camelia, jsaez@eio.uva.es

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_51,

Next we apply Lagrangian Relaxation to the CONPMP based on the resolution of subproblems of type PMP. Instead of applying classical procedures such as subgradient optimization [4], we have seen that an extension of the well-known algorithm of Handler and Zang ([2]), developed for the restricted shortest path problem, is more efficient. This method will be denoted by LARAC algorithm as in ([3]). In addition, we apply the local search improvement heuristic to the found solution. The final result is a very efficient heuristic method.

2 Statement and Notation

CONPMP may be stated as follows:

$$\text{Minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} h_i d_{ij} y_{ij} \tag{1}$$

$$\text{subject to} \quad \sum_{j=1}^{n} y_{ij} = 1 \qquad i = 1,...,m \tag{2}$$

$$y_{ij} \leq x_j \qquad i = 1,...,m, j = 1,...,n \tag{3}$$

$$\sum_{j=1}^{n} x_j = P \tag{4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij} y_{ij} \leq \varepsilon \tag{5}$$

$$x_j \in \{0,1\} \qquad j = 1,...,n \tag{6}$$

$$y_{ij} \in \{0,1\} \qquad i = 1,...,m, j = 1,...,n \tag{7}$$

In this statement, m is the number of demand points of a service, n the number of points of possible establishment of service, P the number of service points which will be established, h_i the demand at point i, d_{ij} the distance between the demand point i and service point j, and q_{ij} is equal to 1 if $d_{ij} > d_C$, and equal to 0 in otherwise. $x_j \in \{0,1\}$ is the localization variable, for $j = 1,...,n$, and $y_{ij} \in \{0,1\}$ is the assignment variable, for $i = 1,...,m, j = 1,...,n$. Note that the left hand side of inequality (5) shows the total covered demand at a distance greater than d_C.

If we denote by $\mathcal{F}$ the feasible set given by the restrictions (2)-(4) and (6)-(7); by $f_1(y) = \sum_{i=1}^{m}\sum_{j=1}^{n} h_i d_{ij} y_{ij}$, and by $f_2(y) = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij} y_{ij}$, the problem can be stated as:

$$\begin{aligned} \text{(P)} \quad &\text{Minimize} && f_1(y) \\ &\text{subject to} && (x,y) \in \mathcal{F} \\ & && f_2(y) \leq \varepsilon \end{aligned} \tag{8}$$

For $k = 1,2$ consider two initial PMP:

$$(P_k) \quad \begin{array}{ll} \text{Minimize} & f_k(y) \\ \text{subject to} & (x,y) \in \mathcal{F} \end{array} \tag{9}$$

Let $(x^k, y^k) \in \mathcal{F}$ be the optimal solution of (P_k). Define $\varepsilon_1 = f_2(y^1)$ and $\varepsilon_2 = f_2(y^2)$.

Property. If the feasible set for PMP $\mathcal{F}$ is nonempty, then

1. Problem (P) is feasible if and only if $\varepsilon_2 \leq \varepsilon$.
2. If $\varepsilon \geq \varepsilon_1$, (x^1, y^1) obeys the restriction $f_2(y) \leq \varepsilon$ and is an optimum for (P).

Thus, we have interest on solving (P) only if it is fulfilled $\varepsilon_2 \leq \varepsilon \leq \varepsilon_1$.

3 Local Search Heuristics for the CONPMP

We develop three heuristically procedures based on local search. These procedures are modifications of the usual exchange search for PMP ([7], [8], [6],[5]), taking into account the restriction (5). The neighborhood structure is: for given $(x,y) \in \mathcal{F}$, $N(x,y)$ consists of those $(x',y') \in \mathcal{F}$ such that x' is obtained from x via a swap exchange, and y' by assigning to each demand point i the closest j with $x'_j = 1$.

Construction Heuristic – H1 Algorithm

Step 1. Choose any solution $(x,y) \in \mathcal{F}$.
Step 2. If $f_2(y) \leq \varepsilon$ end, with $(x,y) \in \mathcal{F}$ a feasible solution.
Step 3. Examine $N(x,y)$. If there is no $(x',y') \in N(x,y)$ such that $f_2(y') < f_2(y)$, end without having obtained a feasible solution.
Step 4. Choose $(x',y') \in N(x,y)$ such that $f_2(y') < f_2(y)$, redefine $(x,y) = (x',y')$ and go back to Step 2.

Improvement Heuristic – H2 Algorithm

Step 1. Let $(x,y) \in\in \mathcal{F}$ with $f_2(y) \leq \varepsilon$.
Step 2. Examine the $(x',y') \in N(x,y)$ which satisfy $f_2(y) \leq \varepsilon$. If there is no $(x',y') \in N(x,y)$ which simultaneously satisfies $f_1(y') < f_1(y)$ and $f_2(y') \leq \varepsilon$, end, since we have arrived at a local optimum.
Step 3. Take $(x',y') \in N(x,y)$ such that $f_1(y') < f_1(y)$ and $f_2(y') \leq \varepsilon$, redefine $(x,y) = (x',y')$ and go back to step 2.

GRASP Multistart Heuristic – H3 Algorithm

The H1 (construction) and the H2 (improvement) algorithms can be integrated on a multistart algorithm, which takes into account two important matters: diversification and intensification. We have chosen the GRASP methodology for the CONPMP. The complete algorithm, with K starts, is (*it* is the iteration index):

Initial Step Set $f^* = \infty$, $it = 0$.
General Step While $it \leq K$, **do**:

1. (**iteration**) Set $it = it + 1$.
2. (**GRASP**) Apply a randomized greedy method with objective $f_1(y)$, and let $(x^1, y^1) \in \mathcal{F}$ be the obtained solution. Generally we will not have $f_2(y^1) \leq \varepsilon$.
3. (**feasibility**) Beginning with $(x^1, y^1) \in \mathcal{F}$ apply algorithm $H1$ to find a feasible solution. If one is found, let $(x^2, y^2) \in \mathcal{F}$, with $f_2(y^2) \leq \varepsilon$. If not, exit the iteration and go back to 1.
4. (**improvement**) Beginning with (x^2, y^2), apply algorithm $H2$, and let (x^3, y^3) be the obtained solution.
5. (**refreshing**). If $f_1(y^3) < f^*$, set $f^* = f_1(y^3)$, $(x^*, y^*) = (x^3, y^3)$.

end do

4 Lagrangian Relaxation. LARAC Method

Dualizing (5), with a multiplier $\lambda \geq 0$, we obtain the Lagrangian relaxation

$$L(\lambda) = \min_{(x,y)\in\mathcal{F}} \sum_i \sum_j (h_i d_{ij} + \lambda q_{ij}) y_{ij} - \lambda \varepsilon.$$

The PMP-subproblem to solve is $P(\lambda) : \min_{(x,y)\in\mathcal{F}} \{f_1(y) + \lambda f_2(y)\}, \forall \lambda \geq 0$.

Initial Step

1. Solve problem P_1, which is the P-median problem with objective $f_1(y)$, and let $(\bar{x}, \bar{y})$ be an optimal solution. If $f_2(\bar{y}) \leq \varepsilon$, then $(\bar{x}, \bar{y})$ is an optimal solution of problem (P) and we are done. Otherwise, set $A = (f_1(\bar{y}), f_2(\bar{y}))$.
2. Solve problem P_2, which is the P-median problem with objective $f_2(y)$, and let $(\bar{x}, \bar{y})$ be an optimal solution. If $f_2(\bar{y}) > \varepsilon$, then problem (P) is not feasible, so we stop. Otherwise, take $B = (f_1(\bar{y}), f_2(\bar{y}))$.

General Step Given $A = (a_1, a_2)$, $B = (b_1, b_2)$ with $a_2 > \varepsilon$, $b_2 \leq \varepsilon$, $a_1 < b_1$.

1. Let be $\lambda = (b_1 - a_1)/(a_2 - b_2)$. Solve the P-median problem $P(\lambda)$, and let $(\bar{x}, \bar{y})$ be an optimal solution, with corresponding point $C = (f_1(\bar{y}), f_2(\bar{y}))$.
2. If C is equal to A or B, end. Otherwise,
3. If $f_2(\bar{y}) \leq \varepsilon$ and $C \neq B$, then $f_1(\bar{y}) = c_1 < b_1$, and we set $B = (f_1(\bar{y}), f_2(\bar{y}))$.
4. If $f_2(\bar{y}) > \varepsilon$ and $C \neq A$, then $c_2 = f_2(\bar{y}) < a_2$, and we set $A = (f_1(\bar{y}), f_2(\bar{y}))$.

5 Computational Results

We present the computational results obtained by applying the above procedures to various CONPMP. Moreover, we have solved the problems to optimality using the solver Xpress Optimizer version 20.00. For this, we generated 30 CONPMP as following. In all cases we took $P = 15$, $m = n$ with values 300, 500, integer demands generated uniformly between 10 and 100, and Euclidean distances from coordinates

generated uniformly between 0 and 1000. For each dimension, we generated 5 problems randomly, and in each case we took 3 values of ε, finally got 30 CONPMP. We give some details on how each algorithm was run.

1. XPRESS. It was run with all default options, excepting a time limit of 600 s.
2. GRASP. In the multistart algorithm H3 we used $K = 20$ initializers, and in the GRASP step $RCL = 80$. In Step 3 (H2), we chosen the rule of steepest descent.
3. LARAC. To solve the PMP subproblems, we used a GRASP method with the Resende-Werneck implementation ([6], [5]), with $K = 20$ and $RCL = 80$.

		Xpress		Grasp			Larac		
Problem	ε	z	t	z^*	t	d	z^*	t	d
$s300\_1$	4000	163931.48	103.06	164413.04	2.39	0.29	164903.22	5.55	0.59
$s300\_1$	4200	160614.96	24.70	161029.65	14.81	0.26	160614.96	4.75	0
$s300\_1$	4500	160254.66	90.45	160254.66	28.55	0	160254.66	5.01	0
$s300\_2$	3700	163670.52	73.05	165456.14	1.45	1.09	163670.52	8.72	0
$s300\_2$	4000	161037.46	99.77	161446.04	16.42	0.25	161053.10	8.23	0.01
$s300\_2$	4300	159620.72	97.66	159620.72	25.17	0	159620.72	14.97	0
$s300\_3$	3500	161004.26	92.70	161203.98	2.53	0.12	161510.12	6.19	0.31
$s300\_3$	3800	157483.89	29.31	157483.89	17.20	0	157483.89	8.98	0
$s300\_3$	4200	156207.86	37.63	156280.76	33.00	0.05	157129.88	4.19	0.59
$s300\_4$	3800	154480.03	108.72	154905.37	6.13	0.28	154480.03	16.56	0
$s300\_4$	4000	151374.96	62.75	152858.45	19.38	0.98	151444.59	4.72	0.05
$s300\_4$	4100	150456.02	57.59	151095.88	20.78	0.43	150456.02	9.53	0
$s300\_5$	4000	162789.76	27.61	165418.72	6.25	1.61	163598.73	5.22	0.5
$s300\_5$	4400	160810.80	20.59	161272.58	23.36	0.29	160810.80	92.86	0
$s300\_5$	4800	160532.80	91.95	160724.02	31.13	0.12	160724.02	6.16	0.12
Average			67.83		16.57	0.38		13.44	0.14

Table 1 Computational performance size 300

Tables 1 and 2 contain a summary of the results obtained. In every case, the first column contains the name of the problem, which refers to m and n, and the next column shows the value chosen for ε, always between the limits ε_2 and ε_1. For each of the three methods (XPRESS, GRASP and LARAC) we have the value of the best solution obtained and the time required to calculate it, in seconds, as well as the deviation (percent) $d = 100 \cdot (z^* - z)/z$ from the optimum.

1. For the 30 problems, Xpress attained the optimal value.
2. For the 15 problems having dimension 300 in Table 1, LARAC found a better solution than GRASP in 8 out of the 15 cases, GRASP a better one than LARAC in 3, and in the remaining 4 cases the methods tied.
3. For the problems having dimension 500 in Table 2, GRASP got a better solution than LARAC in 9 cases and tied in 3 cases.
4. Both the average deviation and time on LARAC are lowest than on GRASP.

		Xpress		Grasp			Larac		
Problem	ε	z	t	z^*	t	d	z^*	t	d
s500_1	7800	288154.98	136.89	288154.98	18.34	0	288154.98	57.78	0
s500_1	8200	286027.72	158.73	286027.72	72.69	0	286065.85	21.42	0.01
s500_1	8600	285476.27	112.94	285476.27	105.88	0	285476.27	23.39	0
s500_2	7100	271464.64	295.64	271736.87	52.45	0.10	271464.64	25.69	0
s500_2	7700	268980.30	196.39	269301.21	123.42	0.12	269279.49	21.22	0.11
s500_2	8300	268095.42	134.70	268095.42	144.70	0	268095.42	27.25	0
s500_3	6700	280053.04	448.89	280239.10	12.20	0.06	280442.80	27.42	0.13
s500_3	7300	277784.10	408.36	278335.10	67.77	0.19	278044.29	50.83	0.09
s500_3	7800	276437.36	431.23	277331.21	112.77	0.32	276454.26	22.19	0.01
s500_4	6700	283852.18	279.73	287154.70	8.94	1.16	284594.67	27.06	0.26
s500_4	7700	279590.46	305.91	280439.71	91.42	0.30	280171.34	18.58	0.21
s500_4	8400	279128.88	229.20	279148.89	141.88	0.01	279274.31	19.42	0.05
s500_5	6900	284614.50	224.41	285633.38	13.39	0.36	284614.50	18.41	0
s500_5	7900	279640.02	380.34	279955.48	92.44	0.11	279640.02	19.17	0
s500_5	8500	278417.05	273.51	278479.73	142.58	0.02	278474.90	28.98	0.02
Average			267.29		80.05	0.18		25.58	0.05

Table 2 Computational performance size 500

5. We may conclude LARAC is more efficient than GRASP, both because of the quality of the solutions that it obtains as well as for the running time.

Acknowledgements The authors wish to thank FICO for providing the Xpress Optimizer application, which has been used to obtain the computational results. The first author is supported by Consejería de Educación de la Junta de Castilla y León (Spain) under the project VA056A09 and by Ministerio de Ciencia e Innovación under the project ECO2008-02358.

References

1. M. S. Daski. *Network and Discrete Location. Models, Algorithms and Applications.* John Wiley and Sons, 1995.
2. G. Handler and I. Zang. A dual algorithm for the constrained path problem. *Networks*, 10: 293–310, 1980.
3. A. Jüttner, B. Szviatovszki, I. Mécs, and Z. Rajkó. Lagrange Relaxation Based Method for the QoS Routing Problem. In *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 859–868, 2001.
4. J. M. Mulvey and H. P. Crowder. Cluster analysis: An application of lagrangian relaxation. *Management Science*, 25: 329–340, 1979.
5. M. G. C. Resende and R. F. Werneck. A hybrid heuristc for the p-median problem. *Journal of Heuristics*, 10: 59–88, 2004.
6. M. G. C. Resende and R.F. Werneck. On the implementation of a swap-based local search procedure for the p-median problem. In *Proceedings of the Fifth Workshop on Algorithm Engineering and Experiments*, pages 119–127, 2003.
7. M. B. Teitz and P. Bart. Heuristic methods for estimating the generalized vertex median of a weigthed graph. *Operations Research*, 16: 955–961, 1968.
8. R. Whitaker. A fast algorithm for the greedy interchange of large-scale clustering and median location problems. *INFOR*, 21: 95–108, 1983.

II.3 Stochastic Programming

Chair: Prof. Dr. Rüdiger Schultz (Universität Duisburg-Essen)

Uncertainty is a prevailing issue in a growing number of optimization problems in science, engineering, and economics. Stochastic programming offers flexible methodology for mathematical optimization problems involving uncertain parameters for which probabilistic information is available. This covers model formulation, model analysis, numerical solution methods, and practical implementations. We solicit contributions to original research from this spectrum of topics.

A Continuous-Time Markov Decision Process for Infrastructure Surveillance

Jonathan Ott

Abstract We consider the dynamics of threat of an infrastructure consisting of various sectors. The dynamics are modelled by a continuous-time Markov decision process (CMDP). A key property of the model is that dependencies between sectors are taken into account. The actions are determined by the number of the available security staff. The decision maker has to allocate the security staff to the sectors which leads to a resource allocation problem. The CMDP criterion is to minimize expected discounted total cost over an infinite horizon. This problem is computable only for very small infrastructures. To solve the problem for larger infrastructures, an index for a specific restless bandit model is introduced. The index can be easily implemented. The index is used as a basis of a heuristics which defines a suboptimal decision rule for the surveillance task. By using this heuristics, memory requirements are reduced in order to store the suboptimal decision rule efficiently.

1 Introduction

Infrastructures like train stations and airports are exposed to multiple threats like natural hazards, crime and terrorism. In order to protect these infrastructures, surveillance is the first choice. But instead of increasing the surveillance means which would lead to informational overload, intelligent technical support should be designed that uses already installed sensors. The goal is to develop a decision support system (DSS) which should improve the situation awareness of the decision maker by providing a global view of the current threat situation and by proposing optimal actions so that global threat of the infrastructure is minimized.

To model the threat dynamics, we use a CMDP. As we shall see, we are able to provide appropriate decision support from the definition and from an approximate solution of the CMDP.

Jonathan Ott
Karlsruhe Institute of Technology, Institute for Stochastics, e-mail: j.ott@kit.edu

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 327
DOI 10.1007/978-3-642-20009-0_52,

2 Mathematical Model

We are going to model the dynamics of threat in terms of a finite-state and finite-action CMDP. In the remainder of this section, we are going to introduce a model for the surveillance task.

CMDPs Following [5], a CMDP consists of several components: There is a finite state space S and a finite action space A. The set $D \subset S \times A$ is called the restriction set and $D(s) := \{a : (s,a) \in D\} \neq \emptyset$, $s \in S$, is the set of all admissible actions in state s. After each state transition, the decision maker chooses an action a from the set $D(s)$ when the current state is s. The dynamics of the CMDP are given by a set of transition rates, meaning that the system remains in state s for an $\mathrm{Exp}(\lambda(s,a))$-distributed time, $(s,a) \in D$. After this time, the system jumps to the subsequent state s' with probability $p^a_{ss'}$. As long as the system occupies state s and action a is being executed, the decision maker has to pay a cost rate $C(s,a)$. Furthermore, we need a discount rate $\alpha > 0$.

A given decision rule $\mu : S \to A$ with $\mu(s) \in D(s)$ for all $s \in S$ and a given initial state $x_0 \in S$ together define a stochastic state process $(X_t)_{t \geq 0}$ with values in S. Let F be the set of all decision rules. A classical problem concerning CMDPs is the search for a decision rule $\mu^* \in F$ such that it minimizes the expected discounted total cost over an infinite horizon, given by $v^\mu(s) := E^\mu[\int_0^\infty e^{-\alpha t} C(X_t, \mu(X_t))\, dt \,|\, X_0 = s]$, $s \in S$, over all decision rules $\mu \in F$. According to, e. g., [5], the optimal value function $v^* := \inf_{\mu \in F} v^\mu$ is the unique solution of the Bellman equation

$$v(s) = \min_{a \in D(s)} \left\{ \frac{C(s,a) + \lambda(s,a) \sum_{s' \in S} p^a_{ss'} v(s')}{\lambda(s,a) + \alpha} \right\}, \quad s \in S, \tag{1}$$

and an action $\mu^*(s)$ is optimal in s if and only if it is a minimizing action of (1). In fact, it is optimal over a much broader class of policies. Methods to derive optimal decision rules include Howard's policy improvement algorithm and linear programming. For the details, we refer to [5].

Surveillance Task To model the dynamics of threat of a given infrastructure, we introduce the following parameters: a given infrastructure consists of several sectors forming the set Σ. Furthermore, there are dependencies between the sectors. Let $N \in \{0,1\}^{\Sigma \times \Sigma}$ with $N(\sigma, \sigma^*) = 1$ if σ^* is dependent from σ, meaning that when threat changes in σ threat may change in σ^*, and $N(\sigma, \sigma^*) = 0$ if a change of threat in σ does not influence threat of σ^*. To measure threat, we assume that each sector has a current *threat level* $g \in \{0, \ldots, g_{\max}\}$, giving the state space $S = G^\Sigma$. Hence, a state is a risk map of the infrastructure and gives an overview of the current threat situation of the infrastructure improving the decision maker's situation awareness.

The decision maker is allowed to assign one *elementary action* from the finite set A_0 to every sector. So, the action space is $A = A_0^\Sigma$. Each elementary action $a_0 \in A_0$ is endowed with certain features: it takes an $\mathrm{Exp}(\lambda_{a_0}(\sigma))$-distributed time to accomplish a_0 in σ. When a_0 is accomplished, then the resulting threat level of σ whose current threat level is $s(\sigma)$ will be $s'(\sigma)$ with probability $\Phi_{a_0}^{\sigma, s(\sigma)}(s'(\sigma))$.

Depending on $s(\sigma)$ and the result $s'(\sigma)$, the threat level of any dependent sector σ^* of σ, will change from $s(\sigma^*)$ to $\varphi_{a_0}(s(\sigma), s'(\sigma), s(\sigma^*)) \in G$. Sectors which are not dependent from σ will not be affected by the completion of a_0. Furthermore, executing a_0 costs $c_{a_0} \geq 0$ per unit time. We assume that there is a special elementary action $0 \in A_0$ whose interpretation is "do nothing." In detail, we define $\lambda_0(\sigma) := 0$, $c_0 := 0$, $\Phi_0^{\sigma,g}(g) := 1$ and $\varphi_0(g, g', g^*) := g^*$ for all $\sigma \in \Sigma$ and g, g', $g^* \in G$.

Events which have an influence on the threat levels of the sectors are called *threat events*. The set of threat events which can occur in $\sigma \in \Sigma$ is denoted by $\mathcal{E}(\sigma)$. A threat event $e \in \mathcal{E}(\sigma)$ is associated with several features: if the current threat level of σ is g, then e will occur after an $\mathrm{Exp}(\lambda_e(g))$-distributed time. When e occurs, it leaves σ with the subsequent threat level $\Psi_e(g) \in G$. The threat levels of some dependent sector of σ changes from g to $\psi_e(g) \in G$. The threat levels of independent sectors remain unchanged. When e occurs, it costs $C_e \geq 0$.

In a practical application, it would not be possible to assign an elementary action to every sector. The number of sectors in which elementary actions apart from 0 can be executed is limited by the staff size r which gives the restriction set $D(s) := \{a \in A : \sum_{\sigma \in \Sigma}[1 - \delta_{0a(\sigma)}] \leq r\}$, $s \in S$, where δ is the Kronecker delta. Assuming that all above mentioned exponentially distributed times are independent, we have a CMDP with cost rates $C(s,a) = \sum_{\sigma \in \Sigma}[c_{a(\sigma)} + \sum_{e \in \mathcal{E}(\sigma)} \lambda_e(s(\sigma))C_e]$, $(s,a) \in D$.

The surveillance task is to minimize the resulting expected discounted total cost over an infinite horizon, which means minimizing overall threat of the infrastructure. Theoretically, such a decision rule is assured. But in practice, an optimal decision rule cannot be derived due to the fact that the state space grows exponentially with the number of sectors. An infrastructure with five sectors, $g_{\max} = 4$, three threat events for each sector and $r = 2$ could be solved exactly by linear programming in about 10 minutes using CPLEX 11.2.0. But increasing the size of the infrastructure to six, CPLEX was not able to compute a solution within 48 hours. Therefore, we need approximation methods or heuristics to solve the problem approximately.

Several approximation methods can be found in the literature. An overview is given in [4]. An approximate linear programming approach can be found in [1] and by exploiting the structure of the model, i. e., only some sectors change their threat levels at a time, this approach can be refined similarly to [3]. But the results of this method are not very convincing. So we are looking for a new approach.

Moreover, memory problems arise when storing a decision rule. If a decision rule is stored naively by one ASCII symbol for each elementary action for each sector in a lookup table, the decision rule of an infrastructure with twelve sectors and $g_{\max} = 4$ would need about 2.7 GB of memory.

3 Index Rule Based Heuristics

In this section, we are going to introduce an index for a specific CMDP model. This index will be used to define a heuristics for the surveillance task.

Index Let us consider a resource allocation problem in which the decision maker has n projects all of which are CMDPs of the following type: the state space S and the action space A are finite. There is a passive action $0 \in D(s)$ for every $s \in S$. The cost rates are given by $C(s,a) = C^{\text{state}}(s) + C^{\text{action}}(a)$, $(s,a) \in D$, where $C^{\text{state}}(s)$, $C^{\text{action}}(a) \geq 0$ and $C^{\text{action}}(0) = 0$. For the transition rates, we assume $\lambda(s,a) = \lambda^{\text{state}}(s) + \lambda^{\text{action}}(a)$, $(s,a) \in D$, and $\lambda^{\text{action}}(0) = 0$. We assume that a state transition is triggered by either the current state or the current action. If the system triggers a state transition, then the subsequent state will be given by the transition probabilities $(p_{ss'})$, whereas if the action a triggers the state transition, it will be given by the transition probabilities $(p^a_{ss'})$. For $a \in D(s)$, we define for $v : S \to \mathbb{R}$

$$T_a v(s) := \frac{C^{\text{state}}(s) + C^{\text{action}}(a)}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(a) + \alpha} + \frac{\lambda^{\text{state}}(s) \sum_{s' \in S} p_{ss'} v(s') + \lambda^{\text{action}}(a) \sum_{s' \in S} p^a_{ss'} v(s')}{\lambda^{\text{state}}(s) + \lambda^{\text{action}}(a) + \alpha}, \quad s \in S.$$

Then the Bellman equation for this project is $v(s) = \min_{a \in D(s)} T_a v(s)$, $s \in S$.

Such a resource allocation problem generalizes the restless bandit model of [6] in that there might be more than just one active action. Note that for each project the parameters may be different. Furthermore, the threat model of section 2 is a project of this kind. If the decision maker is able to work on r projects at a time only, the question arises on which projects he or she should work on in order to minimize expected discounted total cost. Similar problems are considered for the multiarmed bandit model in [2], showing that a decision rule following the Gittins index is optimal, and the restless bandit model in [6], defining an index from a Lagrangian approach. The Whittle index has been found to be a very good heuristics. But the computation of the Whittle index of a specific project is often rather tedious if it exists at all due to the issue of indexability.

Lemma 1 *Let (P) be a project of the above type. If v^* is the optimal value function of (P) and μ^* is an optimal decision rule of (P), then*

$$\iota(s) := \lambda^{\text{action}}(\mu^*(s)) \left(T_0 v^*(s) - \sum_{s' \in S} p^{\mu^*(s)}_{ss'} v^*(s') \right) - C^{\text{action}}(\mu^*(s)) \geq 0, \quad s \in S.$$

Proof. From the Bellman equation, it follows $T_{\mu^*(s)} v^*(s) \leq T_0 v^*(s)$, $s \in S$, from which we easily obtain the assertion.

Let us have a look at $\iota(s)$ from a heuristic point of view. If we either increase $\lambda^{\text{action}}(\mu^*(s))$ or if we decrease $C^{\text{action}}(\mu^*(s))$, then $\iota(s)$ increases. In both cases, we would prefer to work on a project in which $\lambda^{\text{action}}(\mu^*(s))$ is higher or in which $C^{\text{action}}(\mu^*(s))$ is lower if the remaining parameters are the same. So, we could use $\iota(s)$ as a heuristic index. A representation of ι which is easier to compute is given in the following proposition.

Proposition 1 *We have $\iota(s) = (\lambda(s, \mu^*(s)) + \alpha)\,(T_0 v^*(s) - v^*(s))$, $s \in S$.*

Proof. This follows from $v^*(s) + \iota(s)/(\lambda^{\text{state}}(s) + \lambda^{\text{action}}(\mu^*(s)) + \alpha) = T_0 v^*(s)$, $s \in S$.

From this proposition, we obtain an interpretation of ι. The first factor essentially gives the rate until a state transition occurs, or in other words, the rate of allocation of one resource working optimal on the respective project. The second factor compares the values of two decision rules in s. The value $T_0 v^*(s)$ can be seen as a one-step approximation of the value function of the decision rule which acts optimally in every state except for s in which 0 is chosen. Here, the value function v^* is used as a first approximation of the value function of this policy. Of course, the value $v^*(s)$ is the value according to the optimal decision rule in s. Together, $\iota(s)$ measures the approximate "benefit" when using action $\mu^*(s)$ in s instead of using 0.

Proposition 2 *For $s \in S$, we have $\iota(s) \geq 0$, and $\iota(s) = 0$ iff 0 is optimal in s.*

Proof. This follows immediately from the representation of ι in Proposition 1.

Therefore, the index ι identifies states in which it is optimal to use the passive action. Note that for any function $\tilde{\iota}(s) := f(s)\,(T_0 v^*(s) - v^*(s))$, $s \in S$, where $f > 0$, the assertions of Proposition 2 hold. So, variants of ι could be easily constructed.

Heuristics Now, we describe an algorithm to solve the surveillance task from section 2 approximately. Let r be the number of the available security staff.

1. Let $m \geq r$ and consider all subinfrastructures of Σ of size m. The data of a subinfrastructure is given by the original infrastructure Σ restricted to the sectors of the respective subinfrastructure.
2. Solve the surveillance task for all subinfrastructures $\tilde{\Sigma} = \{\sigma_1, \ldots, \sigma_m\}$ exactly, which gives optimal value functions $v^*_{\tilde{\Sigma}}$ and optimal decision rules $\mu^*_{\tilde{\Sigma}}$.
3. Let $s \in S$. Compute the indices $\iota_{\tilde{\Sigma}}(s_{\tilde{\Sigma}})$ for all subinfrastructures $\tilde{\Sigma} = \{\sigma_1, \ldots, \sigma_m\}$ of size m, where $s_{\tilde{\Sigma}}$ is the restriction of s to the sectors of $\tilde{\Sigma}$.
4. Let $\Sigma^*(s)$ be some subinfrastructure with maximal index. Define the heuristic decision rule by

$$\mu_{\text{heur}}(s)(\sigma) := \begin{cases} \mu^*_{\Sigma^*(s)}\left(s_{\Sigma^*(s)}\right)(\sigma), & \text{if } \sigma \in \Sigma^*(s) \\ 0, & \text{else} \end{cases}, \quad \sigma \in \Sigma.$$

In short, the heuristics chooses a subinfrastructure which has maximal index for $s \in S$ and assigns the optimal actions of the very subinfrastructure to the original infrastructure and the passive action to the remaining sectors.

In a small numerical study, we have an infrastructure consisting of five sectors so that the surveillance task is exactly computable. Here, we have $A_0 = \{0, 1, 2\}$, where 1 models a camera evaluation and 2 models an inspection walk. For each sector, we have three threat events: destruction, alarm and a "nothing happened" event. In Table 1, results for the heuristics are given. It shows the relative errors according to the optimal value function. It can be seen that the heuristic decision rule is quite good although it is based on the very simple index ι. Moreover, the study shows that the heuristics tends to assign active actions to valuable sectors. Hence, the proposed actions of this heuristics could be used as a suboptimal decision rule for the DSS.

Table 1 Results from a numerical study.

n	r	m	Minimal relative error	Average relative error	Maximal relative error
5	1	1	6.8 %	30.3 %	93.9 %
5	1	2	20.0 %	57.9 %	114.0 %
5	1	3	13.9 %	39.7 %	81.9 %
5	1	4	4.7 %	14.7 %	45.2 %
5	2	2	2.1 %	17.2 %	46.0 %
5	2	3	1.0 %	11.5 %	41.5 %
5	2	4	0.3 %	4.8 %	26.5 %

Using this heuristics, one need not store the whole decision rule. It is sufficient to store the indices and policies for every subinfrastructure. For a twelve-sector infrastructure approximated by five-sector subinfrastructures, the memory needed is about 75 MB.

4 Conclusion

We presented a CMDP model for threat of an infrastructure. Due to the curse of dimensionality, exact solutions are only available for small infrastructures. To solve the problem approximately, we introduced an index which is easy to compute. Using the index as a basis for a heuristics, gives promising results for the surveillance task. The states and the respective suboptimal actions can be used as a basis of a DSS.

Acknowledgements The underlying projects of this article are funded by the Federal Ministry for Education and Research of the Federal Republic of Germany under promotional references 03BAPAC1 and 03GEPAC2. The author is responsible for the content of this article.

References

1. D. P. de Farias and B. van Roy. The Linear Programming Approach to Approximate Dynamic Programming. *Operations Research*, 51(6): 850–865, November-December 2003.
2. J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2): 148–164, 1979.
3. Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient Solution Algorithms for Factored MDPs. *Journal of Artificial Intelligence Research*, 19(1): 399–468, July 2003.
4. Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2007.
5. Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2005.
6. P. Whittle. Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25: 287–298, 1988.

Stochastic Extensions to FlopC++

Christian Wolf, Achim Koberstein, and Tim Hultberg

Abstract We extend the open-source modelling language FlopC++, which is part of the COIN-OR project, to support multi-stage stochastic programs with recourse. We connect the stochastic version of FlopC++ to the existing COIN class stochastic modelling interface (SMI) to provide a direct interface to specialized solution algorithms. The new stochastic version of FlopC++ can be used to specify scenario-based problems and distribution-based problems with independent random variables. A data-driven scenario tree generation method transforms a given scenario fan, a collection of different data paths with specified probabilities, into a scenario tree. We illustrate our extensions by means of a two-stage mixed integer strategic supply chain design problem and a multi-stage investment model.

1 Introduction

Stochastic programming is about optimal decision making under uncertainty [2]. Uncertain parameters are realistic in most practical decision problems [15], so it would be convenient to easily implement those problems. Modelling environments for deterministic mathematical problems are numerous and widely used. This is not the case for modelling environments for stochastic programming. An overview about existing modelling environments as well as detailed information can be found in [15, 9, 12], examples are SPiNE [14], SLP-IOR [9] and STOCHASTICS [15]. Support for modelling of stochastic programs was also added to some commercial

Christian Wolf
University of Paderborn, Paderborn, Germany e-mail: christian.wolf@dsor.de

Achim Koberstein
University of Paderborn, Paderborn, Germany e-mail: koberstein@dsor.de

Tim Hultberg
Eumetsat, Darmstadt e-mail: tim.hultberg@eumetstat.int

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 333
DOI 10.1007/978-3-642-20009-0_53,

modelling software, e.g. GAMS, MPL, AIMMS, Xpress or Microsofts Solver Foundation.

The process of creating a solvable stochastic program instance is threefold. In the first step, the stochastic program can be implemented using a modelling language designed for that purpose. Algebraic modelling languages (AML) allow for models that are easier to build, read and thus maintain than the use of problem specific matrix generators [5].

A major difficulty for the modeller is the determination of the stochastic process of the real world and the translation into random variables with a certain distribution or statistical parameters. This leads to the second step which consists of the generation of random variates for all the random variables defined in the first step. The generation and combination process is called *scenario generation* and is an active research area of its own [11]. Kaut and Wallace state that there is no "scenario-generation method that would be best for all possible models". Therefore the scenario generation method has to be chosen with respect to the model. The AML should provide the modeller with the possibility to use his scenario generation method of choice.

In the third step, the stochastic problem defined by the core model and the scenario tree gets solved. Several different approaches are possible to reach this goal. The deterministic equivalent problem can be easily generated given the deterministic core model and the scenario tree. It can be solved with standard techniques. With a rising number of scenarios, this approach soon becomes infeasible, as the problem size grows linearly with the number of scenarios, which grows exponentially with respect to the number of (independent) random variables and the number of outcomes to each random variable. This is the so-called "curse of dimensionality".

Because of that effect decomposition methods were proposed that take advantage of the special structure of stochastic programs with recourse. There are two different types of problem formulations for stochastic programs with respect to the non-anticipativity constraints. The node-based or implicit formulation generates variables for every node in the scenario tree, whereas the scenario-based or explicit formulation introduces n copies of each decision variable for every stage, if n is the number of scenarios. All copies are then forced by the explicit non-anticipativity constraints to take the same value. Depending on the formulation different decomposition methods can be applied [6], so the modelling environment should be able to generate both formulations.

Our goal is to provide an open-source modelling environment which renders it possible to specify multistage stochastic programs with recourse and solve these programs. The environment is based on FlopC++ [8], which is part of the COIN-OR project [13]. The COIN-OR project provides open-source software for the operations research community. The storage of a stochastic program is done via the stochastic modelling interface (Smi), which is also part of COIN-OR. Our contribution is based on previous work in the direction of chaining FlopC++ and Smi together [10], which is a plausible approach, as these two projects are already present in COIN-OR. It is theoretically possible to solve all multistage stochastic programs in the deterministic equivalent form with any Osi capable solver. Furthermore spe-

cialized solver for stochastic programs can be connected, if they implement the stochastic solver interface we propose on the basis of Osi and Smi. We used an implementation of the L-shaped method with first stage integer variables [3] with this interface to solve practical problems. We also extended Smi to handle integer variables, so that we can model stochastic integer programs and solve them in the two-stage case. Furthermore the environment provides methods to compute the stochastic measures EVPI and VSS [2].

2 Stochastic Extensions

In the following, we introduce the new language constructs by means of an example, the well known investment model [2, 12].

```
MP_modell investmentModel(new OsiClpSolverInterface());
MP_data start,goal;
start() = 55; goal() = 80;
int numStage=4;int numScen=8; enum{asset1,asset2,numAssets};
```

Any Osi capable solver can be used to solve the model, it is given in the first line.

```
MP_stage T(numStage);
MP_scenario_set scen(numScen);
MP_set assets(numAssets);
```

A key concept in stochastic programming is the notion of a *stage*. The underlying time process of the model can be divided in stages, where each stage corresponds to a time period where new information becomes available and decisions have to be made after observing the new information. The set `MP_stage` specifies the special stage set, which is mandatory for every stochastic program.

```
double scenarios[numStage-1][numAssets][numScen] =
 {    // stage 2
    {1.25,1.25,1.25,1.25,1.06,1.06,1.06,1.06},//asset1
    {1.14,1.14,1.14,1.14,1.16,1.16,1.16,1.16} //asset2
 },{ // stage 3
    {1.21,1.21,1.07,1.07,1.15,1.15,1.06,1.06},//asset1
    {1.17,1.17,1.12,1.12,1.18,1.18,1.12,1.12} //asset2
 }, {// stage 4
    {1.26,1.07,1.25,1.06,1.05,1.06,1.05,1.06},//asset1
    {1.13,1.14,1.15,1.12,1.17,1.15,1.14,1.12}}};//asset2
MP_random_data returns(&scenarios[0][0][0],T,assets);
```

The random parameter `MP_random_data` is in general an algebraic combination of several other parameters or it refers to a random variable with a specific distribution, as it is the case in the example. A `ScenarioRandomVariable` is implicitly created during the initialization of `returns` to store the values for every scenario.

The expected number of entries is given via the special set `MP_scenario_set`. This is only mandatory if the problem is scenario-based. A `MP_random_data` variable can be used everywhere, where a normal `MP_data` variable has been used before, so it is fairly easy to switch from a deterministic to a stochastic model. One has to replace deterministic data with random data and assign a stage index to all second or later stage variables and constraints. No indexation over specific scenarios is required.

```
MP_variable x(T,assets),wealth(T),shortage(T),overage(T);
MP_constraint initialWealthConstr, returnConstr(T);
MP_constraint allocationConstr(T),goalConstr(T);
```

The variables and constraints of the model get initialized with the above statements.

```
initialWealthConstr() = sum(assets,x(0,assets)) == start();
allocationConstr(T) =sum(assets,x(T,assets))== wealth(T);
returnConstr(T+1) = sum(assets,returns(T+1,assets)*x(T,
   assets))== wealth(T+1);//only valid for stage 2 to 4
goalConstr(T.last()) = wealth(T.last()) == goal() +
                overage(T.last()) - shortage(T.last())
MP_expression valueFunction( -1.3*shortage(T.last()) +
                   1.1*overage(T.last()));
```

The second stage constraints are indexed over the set $T+1$, so the first stage is spared. The object `valueFunction` stores the objective function of the model. The expectation over all scenarios is added automatically by the environment.

```
investmentModel.setObjective( valueFunction );
investmentModel.solve(MP_model::MAXIMIZE);
```

The investment model is solved via the deterministic equivalent with the LP solver *Clp* through the `OsiClpSolverInterface`, but this can be changed in the first code line.

Distribution-based problems can also be specified with FlopC++ with the added `RandomVariable` hierarchy. A discrete continuous distribution with step width 0.5 in the interval $[0,3]$ can be specified with the following code.

```
RandomVariable* rPtr;
rPtr = new DiscreteContinuousRandomVariable(0,3,0.5);
```

Dependencies between different random variables can not be specified at the moment. Also inter-stage dependencies can not be modelled explicitly via random processes. It is however possible to generate variates according to sophisticated random processes with external tools and feed the data directly to a `ScenarioRandomVariable` as it was the case in the example.

Apart from file-based exchange via SMPS or OSiL [4] there is no common interface between modelling languages and solver for stochastic programs [7]. We therefore propose an interface similar to Osi that consists mainly of an Osi solver and a Smi instance that holds the model and the already generated scenario tree. A solver developer who implements this interface uses Smi as the data structure and the given solver to solve the subproblems.

3 Practical Experience

We implemented a two-stage strategic network model with binary variables on the first stage (basically a simplified version of the model given in [1]) in FlopC++. We connected a stochastic solver [3] that uses the integer L-shaped method with FlopC++ using our stochastic solver interface. Figure 1 shows that in this case the use of a specialized solver is advantageous to the use of standard solution techniques by contrasting the solution times of our stochastic solver with the solution times for the deterministic equivalent [1]. Thus the use of a modelling environment for stochastic programming in combination with an interface for stochastic solvers is highly useful, if these solvers are available. If not, the deterministic equivalent can still be generated and solved instead.

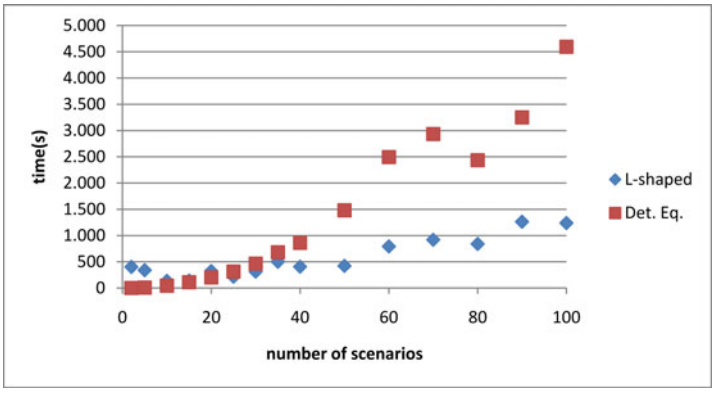


Fig. 1 Comparison of solution time of det. eq. and L-shaped method solver.

[1] All tests have been conducted on a MS Windows XP 64 bit, Intel Xeon 3,4 Ghz, 4 GB RAM with Cplex 11.0. The measured time is the average wall-clock time of three runs.

4 Conclusions and Future Work

We extended the existing open-source AML FlopC++ to support the modelling of multistage stochastic programs with recourse. This modelling framework eases the process of creating stochastic programs opposed to the manual generation of the deterministic equivalent or usage of the column based SMPS format. The environment is restricted to inter-stage independent random variables. In addition we propose an interface for stochastic solvers based on Smi and Osi. An integration of various discretization schemes for continuous random distributions or empirical data within the environment would be valuable [11] and is subject of future research.

References

1. Ralf Bihlmaier, Achim Koberstein, and René Obst. Modeling and optimizing of strategic and tactical production planning in the automotive industry under uncertainty. *OR Spectrum*, 31(2): 311–336, August 2009.
2. John R. Birge and François V. Louveaux. *Introduction to Stochastic Programming*. Springer Verlag, 1997.
3. Corinna Dohle. *Eine Implementierung des Benders-Dekompositionsverfahrens für allgemeine zweistufig stochastische Programme mit diskreten Stufe-1-Variablen.* Diploma thesis, University of Paderborn, April 2010.
4. R. Fourer, H. I. Gassmann, J. Ma, and R. K. Martin. An XML-based schema for stochastic programs. *Annals of Operations Research*, 166(1): 313–337, October 2008.
5. Robert Fourer. Modeling languages versus matrix generators for linear programming. *ACM Transactions on Mathematical Software*, 9(2): 143–183, June 1983.
6. Robert Fourer and L. Lopes. A management system for decompositions in stochastic programming. *Annals of Operations Research*, 142(1): 99–118, 2006.
7. Emmanuel Fragniere and Jacek Gondzio. Stochastic programming from modeling languages. In Stein W. Wallace and William T. and Ziemba, editors, *Applications of Stochastic Programming*, chapter 7, pages 95–113. SIAM, 2005.
8. Tim Helge Hultberg. FLOPC++ An Algebraic Modeling Language Embedded in C++. In *Operations Research Proceedings 2006*, pages 187–190. Springer Berlin-Heidelberg, 2006.
9. Peter Kall and Janos Mayer. *Stochastic linear programming: models, theory, and computation*. Springer, 2005.
10. Michael Kaut, Alan J King, and Tim Helge Hultberg. A C++ Modelling Environment for Stochastic Programming. Technical Report rc24662, IBM Watson Research Center, 2008.
11. Michael Kaut and Stein W. Wallace. Evaluation of scenario-generation methods for stochastic programming. *Stochastic Programming E-Print Series*, 14(1), 2003.
12. Miloš Kopa, editor. *On Selected Software for Stochastic Programming*. 2008.
13. Robin Lougee-Heimer. The Common Optimization Interface for Operations Research. *IBM Journal of Research and Development*, 47(1): 57–66, January 2003.
14. Christian Valente, Gautam Mitra, Mustapha Sadki, and Robert Fourer. Extending algebraic modelling languages for Stochastic Programming. *INFORMS Journal on Computing*, 21(1): 107–122, 2009.
15. S. W. Wallace and W. T. Ziemba, editors. *Applications of stochastic programming*. Society for Industrial Mathematics, 2005.

II.4 Continuous Optimization

Chair: Prof. Dr. Oliver Stein (Karlsruhe Institute of Technology)

Continuous decision variables arise in various practical applications of optimization, ranging from parameter fitting and geometrical problems to questions of sensitivity analysis and robustness.

The theory and methodology of continuous optimization provides a comprehensive framework to handle such problems. In many other disciplines of operations research, like mixed-integer programming, multi-objective programming, and dynamic programming, the results of continuous optimization provide fundamental contributions.

We welcome submissions dealing with original research in theoretical foundations, numerical solution methods, and practical applications of continuous optimization.

Quadratic Order Optimality Conditions for Extremals Completely Singular in Part of Controls

A. V. Dmitruk

Abstract We consider an optimal control problem which is linear in one part of control variables and nonlinear in the state and another part of control variables. For an extremal in this problem, we propose necessary and sufficient optimality conditions of some quadratic order.

1 Introduction

On a fixed time interval $[0,T]$ consider the following optimal control problem:

$$\begin{cases} \dot{x} = f(t,x,u) + F(t,x)\,v, & (1)\\ \eta_j(p) = 0, \quad j = 1,\dots,\mu, & (2)\\ \varphi_i(p) \le 0, \quad i = 1,\dots,\nu, & (3)\\ J = \varphi_0(p) \to \min, & (4) \end{cases}$$

where $x \in \mathbf{R}^n$, $u \in \mathbf{R}^{r_u}$, $v \in \mathbf{R}^{r_v}$, $p = (x(t_0), x(t_1)) \in \mathbf{R}^{2n}$, the functions η_j, φ_i are twice continuously differentiable, f is measurable in t and twice continuously differentiable in x,u, and F is Lipschitz continuous in t and twice continuously differentiable in x. Note that F does not depend on the control u. We call system (1) completely linear in control v.

Let a process $w^0 = (x^0(t), u^0(t), v^0(t))$ satisfy the first order necessary conditions for a weak minimum, i.e., $\exists\,(\alpha_0,\dots,\alpha_\nu) \ge 0$, $(\beta_1,\dots,\beta_\mu)$ with $\sum \alpha_i + \sum |\beta_j| = 1$, and a Lipschitz continuous $n-$ vector function $\psi(t)$ such that, introducing the Pontryagin function $H = \psi\,(f(t,x,u) + F(t,x)\,v)$, the terminal Lagrange

A. V. Dmitruk
Russian Academy of Sciences, CEMI, e-mail: dmitruk@member.ams.org

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_54,

function $l(p)=\sum_{i=0}^{\nu}\alpha_i\varphi_i(p)+\sum_{j=1}^{\mu}\beta_j\eta_j(p)$, and assuming all the inequality constraints to be active, the following relations hold:

i) normalization $\sum\alpha_i+\sum|\beta_j|=1$,

ii) adjoint (costate) equation $-\dot\psi=H_x$,

iii) transversality $\psi(0)=l_{x_0}$, $\psi(T)=l_{x_T}$,

iv) stationarity of H in u: $H_u = \psi(t)f_u(t,x^0(t),u^0(t)) \equiv 0$,

v) stationarity of H in v:

$$H_v = \psi(t)F(t,x^0(t)) \equiv 0 \qquad \text{(partially singular extremal)}.$$

The set of these relations is called Euler–Lagrange equation.

For simplicity in presentation, assume here that the collection $(\alpha,\beta,\psi(t))$ is unique, up to normalization.

Here we propose higher order conditions for the two types of minimum at w^0: 1) the weak minimum with respect to both u and v (weak–weak minimum), and 2) the weak minimum with respect to u and Pontryagin minimum with respect to v (weak–Pontryagin minimum). The latter notion means that, for any number N, the process $w^0=(x^0,u^0,v^0)$ is a local minimum w.r.t. the norm $C\times L_\infty\times L_1$ under the additional constraint $||v-v^0||_\infty\le N$ (see [3, 4]).

The case of Pontryagin minimum in u and weak minimum in v was considered in [2]. If the control u is absent in (1), the second order conditions both for the weak and Pontryagin minimum in v were given in [1, 3].

Compose the Lagrange function

$$\Phi(w) = l(x_0,x_T)+\int_0^T \psi\,(\dot x - f(t,x,u)-F(t,x)\,v)\,dt,$$

and calculate its second variation at w^0:

$$\Omega(\bar w) = (l''_{x_0x_0}\bar x_0,\bar x_0)+2(l''_{x_Tx_0}\bar x_0,\bar x_T)+(l''_{x_Tx_T}\bar x_T,\bar x_T)-$$

$$-\int_0^T\left((H_{xx}\bar x,\bar x)\ +\ 2(\bar x,H_{xu}\bar u)\ +\ (H_{uu}\bar u,\bar u)\ +\ 2(\bar x,H_{xu}\bar v)\right)dt.$$

Define the cone $\mathcal{K}$ of critical variations $\bar w=(\bar x,\bar u,\bar v)$:

$$\dot{\bar x} = (f'_x\bar x+F'_x\bar x v^0)+f'_u\bar u+F\,\bar v, \tag{5}$$

$$\eta'_{x_0}\bar x_0+\eta'_{x_T}\bar x_T\ =0,$$

$$\varphi'_{x_0}\bar x_0+\varphi'_{x_T}\bar x_T\ \le 0,$$

where $\varphi=(\varphi_0,\dots,\varphi_\nu)$. Define the so-called violation function

$$\sigma(w) = \sum\varphi_i^+(x_0,x_T)+|\eta(x_0,x_T)|+\int_0^T|\dot x-f(t,x,u)-F(t,x)v|\,dt.$$

Define the following quadratic functional, that we regard as a quadratic order of minimum:

$$\gamma(\bar{x},\bar{u},\bar{v}) = |\bar{x}(0)|^2 + \int_0^T (|\bar{u}|^2 + |\bar{y}|^2)\,dt + |\bar{y}(T)|^2,$$

$$\text{where} \qquad \dot{\bar{y}} = \bar{v}, \qquad \bar{y}(0) = 0.$$

(Here $\bar{y}$ is an additional, artificial state variable).

2 The Weak–Weak Minimum

Theorem 1. (Conditions for weak–weak minimality)
a) If w^0 is a weak–weak minimum, then

$$\Omega(\bar{w}) \geq 0 \qquad \forall \bar{w} \in \mathcal{K}. \tag{6}$$

b) If $\exists a > 0$ such that

$$\Omega(\bar{w}) \;\geq\; a\,\gamma(\bar{w}) \qquad \forall \bar{w} \in \mathcal{K}, \tag{7}$$

then w^0 is a strict weak–weak minimum. Moreover, $\exists C$ and $\varepsilon > 0$ such that $\sigma(w) \geq C\gamma(w - w^0)$ in the $\varepsilon-$ neighborhood of w^0 w.r.t. $C \times L_\infty \times L_\infty -$ norm.

We can also give other necessary conditions. Rewrite equation (5) in the form

$$\dot{\bar{x}} = A(t)\bar{x} + B(t)\bar{u} + D(t)\bar{v},$$

where

$$A(t) = f_x'(t,x^0,u^0) + F_x'(t,x^0,u^0)\,v^0(t), \;\; B(t) = f_u'(t,x^0,u^0), \;\; D(t) = F(t,x^0(t)).$$

Observing that Ω is degenerate in v, perform the so-called Goh transformation: change the state $\bar{x} \longmapsto (\bar{\xi},\bar{y})$, where $\bar{x} = \bar{\xi} + D\bar{y}$. Then we obtain

$$\dot{\bar{\xi}} = A(t)\,\bar{\xi} + B(t)\bar{u} + D_1\,\bar{y}, \qquad \bar{\xi}(0) = \bar{x}(0), \tag{8}$$

where $D_1 = AD - \dot{D}$. The quadratic form now looks as follows:

$$\Omega(\bar{w}) = b(\bar{\xi}_0, \bar{\xi}_T, \bar{y}_T) +$$

$$+ \int_0^T ((G\bar{\xi},\bar{\xi}) + 2(P\bar{\xi},\bar{u}) + (R\,\bar{u},\bar{u}) + (Q\,\bar{y},\bar{y}) + (V\,\bar{y},\bar{v}))\,dt,$$

where b is a finite-dimensional quadratic form, the matrices G,P,R,Q are measurable and bounded, and V, being a skew-symmetric part of $H_{vx}F$, is Lipschitz

continuous. (The symmetric part and the term $(C\bar\xi,\bar v)$ are integrated by parts.) We do not give here the precise expressions of these matrices.

If $\Omega \ge 0$, it must satisfy the Legendre condition $R(t) \ge 0$ and the Goh conditions (see [4, 7]):

$$V(t) \equiv 0, \qquad Q(t) \ge 0.$$

Then $\bar v$ disappears from Ω, so, taking into account some density arguments (see details in [7]), we can ignore the relation $\dot{\bar y} = \bar v$ and consider $\bar y \in L_2[0,T]$ as a new control variable. The terminal value $\bar y(T)$ should be replaced by an additional parameter $\bar h \in \mathbf{R}^{r_v}$.

Thus we obtain an almost classical quadratic form

$$\Omega(\bar w) = S(\bar\xi_0,\bar\xi_T,\bar h) + \int_0^T \big((G\bar\xi,\bar\xi) + 2(P\bar\xi,\bar u) + (R\bar u,\bar u) + (Q\bar y,\bar y)\big)\,dt,$$

which should be considered on the set of quadruples $(\bar\xi,\bar u,\bar y,\bar h)$ satisfying eq. (8) and the relations:

$$\begin{aligned} &\eta'_{x_0}\,\bar\xi(0) + \eta'_{x_T}\,(\bar\xi(T) + D(T)\bar h) = 0,\\ &\varphi'_{x_0}\,\bar\xi(0) + \varphi'_{x_T}\,(\bar\xi(T) + D(T)\bar h) \le 0. \end{aligned} \tag{9}$$

If Ω satisfies the strong Legendre condition

$$R(t) \ge a > 0, \qquad Q(t) \ge a > 0 \qquad \text{a.e. on } [0,T],$$

we can determine its "sign" by finding the conjugate point, which can be performed, in the case when all multipliers $\alpha_i > 0$ (and so, all inequalities (9) can be taken in the equality form), by a slight modification of the classical procedure of solving the corresponding Euler–Jacobi equation [6].

Note that now γ is nothing but the square of the norm in the Hilbert space of those quadruples.

3 The Weak–Pontryagin Minimum

Here the reference process w^0 must satisfy some additional conditions.

Define the cubic functional

$$\rho(\bar w) = \int_0^T \left[\frac{1}{2}\,(H_{vxx}\bar x,\bar x,\bar v) + ((F_x\bar x,\bar v),H_{xv}\bar y)\right] dt,$$

which, as was shown in [3], is the most essential part of the *third variation* of Lagrange function, and substituting $\bar x \mapsto D\bar y$, obtain the cubic functional

$$e(\bar y) = \int_0^T (\mathcal{E}\bar y,\bar y,\bar v)\,dt,$$

where $\mathcal{E}(t) = \{\mathcal{E}_{ijk}(t)\}$ is a third rank tensor with Lipschitz continuous entries.

Further, for any fixed t_* introduce the differential 1-form

$$\omega(t_*) = (\mathcal{E}(t_*)\,\bar{y},\bar{y},d\bar{y}) = \sum \mathcal{E}_{ijk}(t_*)\,\bar{y}^i\,\bar{y}^j\,d\bar{y}^k\,.$$

Theorem 2. (Conditions for weak–Pontryagin minimality)
a) If w^0 is a weak–Pontryagin minimum, then for any t_* the differential 1-form $\omega(t_*)$ must be closed:

$$d\omega(t_*) = \sum \mathcal{E}_{ijk}(t_*)\,(\bar{y}^i\,d\bar{y}^j + \bar{y}^j\,d\bar{y}^i)\wedge d\bar{y}^k = 0, \tag{10}$$

and the second variation must be nonnegative:

$$\Omega(\bar{w}) \geq 0 \qquad \forall\,\bar{w}\in\mathcal{K}. \tag{11}$$

b) If (7) holds and $\exists\, a>0$ such that

$$\Omega(\bar{w}) \;\geq\; a\,\gamma(\bar{w}) \qquad \forall\,\bar{w}\in\mathcal{K}, \tag{12}$$

then w^0 is a strict weak–Pontryagin minimum. Moreover, $\exists C$ such that $\forall N$ $\exists\,\varepsilon>0$ such that $\sigma(w)\geq C\gamma(w-w^0)$ on the set

$$||x-x^0||_C<\varepsilon, \qquad ||u-u^0||_\infty<\varepsilon,$$
$$||v-v^0||_\infty\leq N, \qquad ||v-v^0||_1<\varepsilon.$$

Relation (10) is an additional condition of equality type, by which the conditions for the weak-Pontryagin minimum differ from those for the weak–weak minimum.

Theorems 1 and 2 are new. The proofs are combinations of those in [3] and [5]. Moreover, similar results are valid if the Lagrange multipliers are not unique.

4 A Special Case: $\dim v = 1$

Here the system (1) can be written as $\dot{x} = f(t,x,u) + v\,g(t,x)$.

In this case we can also consider the weak minimum in u and *bounded-strong* minimum in v (see [5]), which means that $\forall N$ the triple $w^0=(x^0,u^0,v^0)$ is a local minimum w.r.t. the norm

$$|x(0)| + ||y||_C + ||u||_\infty \tag{13}$$

on the set $||v||_\infty\leq N$, where y is an additional state variable satisfying the equation $\dot{y}=v$, $y(0)=0$. (In a sense, $y(t)$ is the principal state variable).

In other words, $\forall N \ \exists \varepsilon > 0$ such that w^0 is a minimum point on the set

$$\begin{aligned} &|x(0)-x^0(0)| < \varepsilon, \qquad &&||y-y^0||_C < \varepsilon, \\ &||u-u^0||_\infty < \varepsilon, \qquad &&||v-v^0||_\infty \le N. \end{aligned} \tag{14}$$

Obviously, this minimum is stronger than the weak–Pontryagin minimum.

Theorem 3. (Sufficient conditions for weak– bounded-strong minimality). *If* $\dim v = 1$, *and* $\exists a > 0$ *such that*

$$\Omega(\bar w) \ \ge \ a\,\gamma(\bar w) \qquad \forall\, \bar w \in \mathcal{K}, \tag{15}$$

then w^0 *is a strict weak – bounded-strong minimum. Moreover,* $\exists C$ *such that* $\forall N \ \exists \varepsilon > 0$ *such that* $\sigma(w) \ge C\gamma(w-w^0)$ *on the set (14).*

Remark. If $g(t,x^0(t)) \ne 0$ for all t, then the minimality w.r.t. the norm (13) is equivalent to that w.r.t. the semi-norm $||x||_C + ||u||_\infty$, i.e. one can replace $|x(0)| + ||y||_C$ by the usual $||x||_C$. This is due to

Lemma 1. *Let* $g(t,x^0(t)) \ne 0$ *for all* t, *and a sequence* $w_k = (x_k, u_k, v_k)$ *be such that* $||x_k - x^0||_C \to 0, \quad ||u_k - u^0||_C \to 0, \quad ||v_k||_\infty \le const$. *Then also* $||y_k - y^0||_C \to 0$.

Note also the following property.

Lemma 2. *If* $g = g(x)$ *does not depend on* t, *and (12) holds, then definitely* $g(x^0(t)) \ne 0$ *for all* t, *and hence the concepts of strong minima with respect to* x *and* y *coincide.*

References

1. A.V. Dmitruk. Quadratic Conditions for a Weak Minimum for Singular Regimes in Optimal Control Problems. *Soviet Math. Doklady*, v. 18, no. 2, 1977.
2. V.A. Dubovitskij. Necessary and Sufficient Conditions for a Pontryagin Minimum in Problems of Optimal Control with Singular Regimes and Generalized Controls (in Russian). *Uspekhi Mat. Nauk*, v. 37, no. 3, p. 185–186, 1982.
3. A.V. Dmitruk. Quadratic Conditions for a Pontryagin Minimum in an Optimal Control Problem, Linear in the Control. *Math. of USSR, Izvestija*, v. 28, no. 2, 1987.
4. A.V. Dmitruk. Quadratic Order Conditions of a Local Minimum for Singular Extremals in a General Optimal Control Problem. *Proc. Symposia in Pure Math., v. 64 "Diff. Geometry and Control" (G.Ferreyra et al., eds.)*, American Math. Society, p. 163–198, 1999.
5. A.A. Milyutin and N.P. Osmolovskii. *Calculus of Variations and Optimal Control*. American Math. Society, Providence, RI, 1999.
6. A.V. Dmitruk. Jacobi Type Conditions for Singular Extremals. *Control and Cybernetics*, v. 37, no. 2, p. 285–306, 2008.
7. A.V. Dmitruk and K. K. Shishov. Analysis of a Quadratic Functional with a Partly Singular Legendre Condition. *Moscow University Comput. Mathematics and Cybernetics*, v. 34, no. 1, p. 16–25, 2010.

On the Scalarization of Set-Valued Optimization Problems with Respect to Total Ordering Cones

Mahide Küçük, Mustafa Soyertem, and Yalçın Küçük

Abstract A construction method of total ordering cone on n dimensional Euclidean space was given, it was shown that any total ordering cone is isomorphic to the lexicographic cone, also, existence of a total ordering cone that contain given cone with a compact base was shown and by using this cone, a solving method of vector and set-valued optimization problems was given recently by Küçük et.al. In this work, it is shown that the minimal element for the set-valued optimization problem with respect to the total ordering cone is also minimal element for the corresponding optimization problem. In addition, we give examples that show the absence of the relationships between the continuity of a set valued map and K-minimal element of this map.

1 Introduction

The main purpose of vector optimization problems is to find optimal elements of a given set in partially ordered linear spaces. Set valued optimization is an extension of vector optimization. Methods for solving set-valued optimization problems are similar to the vector optimization methods. Set valued optimization problems has received increasing attention in recent decades. Recent developments on vector and set-valued optimization can be found in [1, 2, 3, 4, 5]

Scalarization methods convert vector or set-valued problems into real valued problems. Scalarization is used for finding optimal solutions of vector valued optimization problems in partially ordered spaces.

Mahide Küçük
Anadolu University, Yunus Emre Campus Eskişehir, Turkey, e-mail: mkucuk@anadolu.edu.tr

Mustafa Soyertem
Anadolu University, Yunus Emre Campus Eskişehir, Turkey, e-mail: soyertem@gmail.com

Yalçın Küçük
Anadolu University, Yunus Emre Campus Eskişehir, Turkey, e-mail: ykucuk@anadolu.edu.tr

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_55,

A construction method of an orthogonal base of $\mathbb{R}^n$ and total ordering cones on $\mathbb{R}^n$ using any non-zero vector in $\mathbb{R}^n$ was given in [6]. In addition, a solving method for vector and set-valued optimization problems with respect to a total ordering cone was given by using scalarization.

Following this introduction, we first provide the necessary basic definitions. We show the existence of vector valued function derived from a given set-valued mapping. By using this vector-valued function, we also show that the minimal element for the vector optimization problem with respect to the total ordering cone is the minimal element of the given set-valued optimization problem. Finally, we show continuity relationships between set-valued maps and vector valued maps derived from this set-valued maps.

2 Mathematical Preliminaries

In this section, we provide some basic notations, definitions and propositions.

For a vector space Y, any binary relation which have the properties reflexivity, anti-symmetricalness, transitivity, compatibility with addition and compatibility with scalar multiplication is called a partial order on Y. For a pointed (i.e., $C \cap (-C) = \{0\}$), convex cone $C \subset Y$ the relation defined by $y_1 \leq_C y_2 \Leftrightarrow y_2 \in y_1 + C$, $\forall y_1, y_2 \in Y$ is a partial order on Y (Proofs and examples can be found in [10]).

Moreover, if any partial order on Y compares any two vector on Y then, it is called a total order. In this study, we mainly use total orders. So, we need some important properties of them (for proofs of the following properties and additional information [6]). If a pointed, convex ordering cone K satisfies $K \cup (-K) = Y$ then "$\leq_K$" is a total order on Y. For an ordered orthogonal set $\{r_1, r_2, ..., r_n\} \subset \mathbb{R}^n$, the set $K = \left[\bigcup_{i=1}^{n} \{r \in \mathbb{R}^n : \forall j < i, \langle r_j, r \rangle = 0, \langle r_i, r \rangle > 0\} \right] \bigcup \{0\}$ is a total ordering cone on $\mathbb{R}^n$. Every total order on $\mathbb{R}^n$ can be represented by such a cone.

Let C be a cone and $0 \notin B \subset C$. If for all $c \in C$ there exists unique $b \in B$ and $\lambda > 0$ such that $c = \lambda b$ then B is said to be a base of cone C. In addition, if B is compact then, it is said that C has a compact base.

The following theorem shows the relationship between total ordering cones and the cones with a compact base.

Theorem 1 *Let C be a cone in $\mathbb{R}^n$. If C has a compact base then there is a total ordering cone K such that $C \setminus \{0\} \subset int(K)$.*

The next two theorems give the properties of the minimal element of a set with respect to a total ordering cone. The first theorem gives the uniqueness of the minimal element and the second one gives the relationship between minimality and strong minimality.

Let C be an ordering cone, A be a set and $\bar{x} \in A$. If $A \subset \bar{x} + C$ then $\bar{x}$ is called a strongly minimal element of A with respect to cone C.

Theorem 2 *Let K be a total ordering cone in $\mathbb{R}^n$. If a set $A \subset \mathbb{R}^n$ has a minimal element with respect to this cone then this minimal element is unique.*

Theorem 3 *Let $K \subset \mathbb{R}^n$ be a total ordering cone, $A \subset \mathbb{R}^n$ and $\bar{x} \in A$. $\bar{x}$ is a minimal element of A with respect to K if and only if $\bar{x}$ is a strongly minimal element of A with respect to K.*

3 Vectorization

In this section, we get vector valued functions from a special class of set-valued maps. Moreover, we find some relationships between the set-valued maps and the vector valued functions we derived from the set-valued maps. The following theorem gives the existence of the vector valued function. Let C be an ordering cone and $A \subset Y$. If $A+C$ is closed then A is said to be C-closed, if there exists $y \in Y$ such that $A \subset y+C$ then A is said to be C-bounded.

Theorem 4 *Let X be a nonempty set and $C \subset \mathbb{R}^n$ be a cone with a compact base and $int(C) \neq \emptyset$. If $F : X \rightrightarrows \mathbb{R}^n$ is C-closed, C-bounded set valued map then, there exists a vector valued function $V_F : X \to \mathbb{R}^n$ and there exists a total ordering cone K such that $\{V_F(x)\} = min(F(x), K)$, $\forall x \in X$.*

Proof. By Proposition 1.10 [7] the cone C has a compact base then there is an orthogonal set $\{r_1, ..., r_n\}$ and a total ordering cone K such that

$$K = \left(\bigcup_{i=1}^{n} \{a \in \mathbb{R}^n : \forall j < i, \langle r_j, a \rangle = 0, \langle r_i, a \rangle > 0\} \right) \cup \{0\} \tag{1}$$

where $C \subset \{r \in \mathbb{R}^n : \langle r_1, r \rangle > 0\} \cup \{0\}$ and $B := \{c \in C : \langle r_1, r \rangle = 1\}$ is a compact base of C.

Since $F(x)$ is C-bounded, for each $x \in X$ there is a $y \in \mathbb{R}^n$ such that $F(x) \subset \{y\} + C$. (i.e. $y \leq_C \tilde{y}$, for all $\tilde{y} \in F(x)$). Because of $r_1 \in C^\sharp$, $\langle r_1, \cdot \rangle$ is strictly increasing with respect to the cone C (the scalarization section in [8]). So, $\langle r_1, y \rangle \leq \langle r_1, \tilde{y} \rangle$, for all $\tilde{y} \in F(x)$. Therefore, the set $\{\langle r_1, \tilde{y} \rangle : \tilde{y} \in F(x)\}$ is bounded from below.

Since the set of minimal elements of $F(x)$ is also the set of minimal elements of $F(x)+C$, $F(x)+C$ is closed and bounded from below, let $a := min\{\langle r_1, \tilde{y} \rangle : \tilde{y} \in F(x)\}$ and $b := \langle r_1, y \rangle$. It is obvious that $b \leq a$.

The set of minimal elements of the scalar problem

$$(SP_1) \begin{cases} min \langle r_1, \tilde{y} \rangle \\ \tilde{y} \in F(x) \end{cases} \tag{2}$$

is in the form of $F(x) \cap ((a-b) \cdot B + \{y\})$ (Figure 1)

Because the set of minimal elements of $F(x)$ and $F(x)+C$ are the same, we get

$$(F(x)+C)\cap((a-b)\cdot B+\{y\})=F(x)\cap((a-b)\cdot B+\{y\})$$

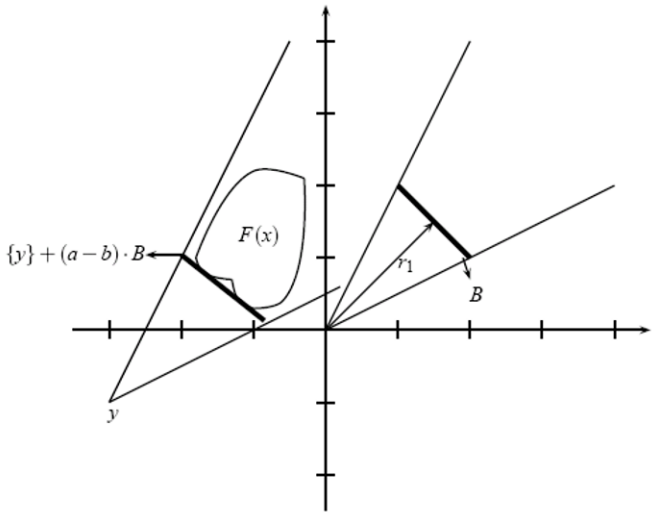


Fig. 1 The set of minimal elements of (SP_1)

Since $F(x)+C$ is closed and $((a-b)\cdot B+\{y\})$ is compact then the set $A_1 := F(x)\cap((a-b)\cdot B+\{y\})$ is also compact.

The set of minimal elements of the scalar problem

$$(SP_2)\begin{cases} min\langle r_2,\tilde{y}\rangle \\ \tilde{y}\in A_1 \end{cases} \tag{3}$$

is in the form $A_2 := A_1\cap\{r\in\mathbb{R}^n : \langle r_2,r\rangle = c\}$ for a $c\in\mathbb{R}$. Since A_1 is compact and the hyperplane $\{r\in\mathbb{R}^n : \langle r_2,r\rangle = c\}$ is closed A_2 is nonempty and compact.

If we go on we get the scalar problem

$$(SP_i)\begin{cases} min\langle r_i,\tilde{y}\rangle \\ \tilde{y}\in A_{i-1} \end{cases} \tag{4}$$

for each $i\in\{3,...,n\}$. And the set of minimal elements of these problems are nonempty and compact.

Since K is a total ordering cone, the last set of minimal elements A_n has only one element. If we define $\forall x\in X$, $\{V_F(x)\} = A_n = \min(A_n,K) = min(F(x),K)$ then $V_F : X\to\mathbb{R}^n$ is a vector valued function.

The following definition, remark end lemma were given in [9].

Definition 1 *Let $C\subset\mathbb{R}^n$ be an ordering cone and A,B any two C-bounded and C-closed nonempty subsets of $\mathbb{R}^n$.*

Then

$$A \leq_C B \Leftrightarrow B \subset A + C$$

Remark 1 *The relation given in Definition 1 for cone bounded and cone closed sets, is reflexive, transitive, compatible with vector addition and scalar multiplication. But it is not anti-symmetric.*

Lemma 1 *Let relation "* $\leq_K$ *" be as in Definition 1 for a total ordering cone* K *and* A, B *be two* C*-bounded and* C*-closed nonempty subsets of* $\mathbb{R}^n$*. Then* $A \leq_K B$ *or* $B \leq_K A$*.*

In the following Theorem we give the representative property of K-minimal elements of sets.

Theorem 5 *Let* $C \subset \mathbb{R}^n$ *be a cone with a compact base and* $int(C) \neq \emptyset$ *and* $F \subset \mathbb{R}^n$ *be a* C*-closed,* C*-bounded set. Let* $V_F \in F$ *be the* K*-minimal of* F *(as we get in Theorem 4, where* K *is a total ordering cone containing the cone* C *as in Theorem 1). Then* $\{V_F\} + K = F + K$*.*

The following corollary gives ordering property of K-minimal elements.

Corollary 1 *Let* $C \subset \mathbb{R}^n$ *be a cone with a compact base and* $int(C) \neq \emptyset$*. Let* $F_1, F_2 \subset \mathbb{R}^n$ *be* C*-closed,* C*-bounded sets and* V_{F_1}, V_{F_2} *be the* K*-minimal elements of* F_1 *and* F_2 *respectively (as we get in Theorem 4) where* K *is a total ordering cone in (1). Then,*

$$F_1 \leq_K F_2 \Leftrightarrow V_{F_1} \leq_K V_{F_2}$$

The following corollary shows that the solution of the vector-valued problem is the solution of set-valued problem with respect to the total ordering cone.

Corollary 2 *Let* X *be a nonempty set and* $C \subset \mathbb{R}^n$ *be a cone with a compact base and* $int(C) \neq \emptyset$*. And,* $F : X \rightrightarrows \mathbb{R}^n$ C*-closed,* C*-bounded set valued mapping.* $V_F : X \to \mathbb{R}^n$ *be vector valued function and* K *is the total ordering cone in (1).*

Then the solution of the set valued problem

$$(SP) \begin{cases} min\ F(x) \\ s.t.\ x \in X \end{cases} \tag{5}$$

with respect to the total order cone K *is same with the solution of the vector problem*

$$(VP) \begin{cases} min\ V_F(x) \\ s.t.\ x \in X \end{cases} \tag{6}$$

The set valued mapping and the vector valued function related with the given set valued mapping have many properties. But continuity is not one of them. The first example shows that the continuity of set valued mapping does not imply the continuity of vector valued function. And the second example shows that also the continuity of vector valued function related with the given set valued mapping does not imply the continuity of set valued mapping.

Example 1 *Let* $F:[0,2\pi)\rightrightarrows\mathbb{R}^2$, $F(x)=[(0,0),(cosx,sinx)]$ *be a set valued map,* $C=\mathbb{R}^2_+$ *and let* K *be the total ordering cone constructed with the orthogonal vectors* $r_1=(1,1), r_2=(-1,1)$*; i.e.*

$$K=\{(x,y)\in\mathbb{R}^2 : x+y>0\}\cup\{(x,y)\in\mathbb{R}^2 : x<0, y=-x\}\cup\{(0,0)\}.$$

Since F *is point-compact set valued map (i.e.* $F(x)$ *is compact for each* $x\in[0,2\pi)$*), it is* C*-bounded and* C*-closed. Moreover, it is continuous with respect to Hausdorff metric. The vector valued function* V_F *derived from* F *with respect to* K *is*

$$V_F(x)=\begin{cases}(cosx,sinx) & : x\in\left(\frac{3\pi}{4},\frac{7\pi}{4}\right]\\ (0,0) & : x\notin\left(\frac{3\pi}{4},\frac{7\pi}{4}\right]\end{cases}. \tag{7}$$

It is obvious that $V_F(x)$ *is not continuous at* $\frac{3\pi}{4}$ *and* $\frac{7\pi}{4}$.

Example 2 *Let* C *and* K *be as in Example 1 and let the set valued mapping* $F:\mathbb{R}\rightrightarrows\mathbb{R}^2$ *be defined as:*

$$F(x)=\begin{cases}[(0,0),(1,2)] : x\in\mathbb{Q}\\ [(0,0),(2,1)] : x\notin\mathbb{Q}\end{cases}. \tag{8}$$

It is obvious that F *is not continuous at any point of* $\mathbb{R}$*. But the vector valued function of* V_F *derived from* F *with respect to* K *is* $V_F(x)=(0,0)$, $\forall x\in\mathbb{R}$ *and* V_F *is continuous on* $\mathbb{R}$.

References

1. Pardalos, P.M.: Pareto Optimality, Game Theory and Equilibria. co-editors: Chinchuluun, A., Pardalos, P.M., Migdalas, A. and Pitsoulis, L. Edward Elgar Publishing, (2008).
2. Chinchuluun, A., Pardalos, P.M.: A survey of recent developments in multiobjective optimization. Annals of Operations Research volume 154, issue 1, pp. 29–50, (2007).
3. Aubin, J.-P.,Cellina, A.: Differential Inclusions. Set-Valued Maps and Viability Theory. Grundlehren Math. Wiss., vol. 264, Springer-Verlag, Berlin (1984).
4. Chen, G.Y., Jahn, J.: Optimality conditions for set-valued optimization problems. Set-valued optimization, Math. Methods Oper. Res. 48, 187–200 (1998).
5. Klein, E., Thompson, A.C.: Theory of correspondences. Including applications to mathematical economics, Canad. Math. Soc. Ser. Monographs Adv. Texts. Wiley and Sons, New York (1984).
6. Küçük, M., Soyertem, M. and Küçük, Y.: On Constructing Total Orders and Solving Vector Optimization Problems with Total Orders. JOGO DOI 10.1007/s10898-010-9576-y, (2010).
7. Luc, D.T.: Theory of Vector Optimization. Springer, Berlin (1989).
8. Jahn, J.: Vector Optimization. Springer, Heidelberg (2004).
9. D. Kuroiwa, Some duality theorems of set-valued optimization with natural criteria, Proceedings of the International Conference on Nonlinear Analysis and Convex Analysis, World Scientific, River Edge, NJ (1999), pp. 221–228.
10. Ehrgott, M.: Multicriteria Optimization. Springer, Berlin (2005).

A Branch & Cut Algorithm to Compute Nondominated Solutions in MOLFP via Reference Points

João Paulo Costa and Maria João Alves

Abstract Based on some previous work on a Branch & Bound algorithm to compute nondominated solutions in multiobjective linear fractional programming (MOLFP) problems using reference points, we now present a new algorithm where special cuts are introduced. These cuts stem from some conditions that identify parts of the feasible region where the nondominated solution that optimizes the current achievement scalarizing function (ASF) cannot be obtained. Introducing the cuts substantially improves the performance of the previous algorithm. The results of several tests carried out to evaluate the performance of the new Branch & Cut algorithm against the old Branch & Bound are reported.

1 Introduction

Reference point techniques provide a useful means for calculating nondominated solutions of MOLFP problems, since they can reach not only supported but also unsupported nondominated solutions (i.e., nondominated solutions that are dominated by unfeasible convex combinations of other nondominated solutions). In fact, it can be observed that the nondominated solution set of a MOLFP problem has, in general, a significant part of unsupported nondominated solutions. The ASF (which temporarily transforms the vector optimization problem into a scalar problem) used by reference point techniques needs to include a sum of objective functions component in order to guarantee that the solutions are nondominated. Otherwise, the ASF only guarantees that the computed solutions are weakly-nondominated. In MOLFP the sum of objective functions leads to a very hard problem to solve (also known as

João Paulo Costa
University of Coimbra/INESC-Coimbra, Portugal, e-mail: jpaulo@fe.uc.pt

Maria João Alves
University of Coimbra/INESC-Coimbra, Portugal, e-mail: mjalves@fe.uc.pt

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_56,

the sum-of-ratios problem) which is essentially NP-hard. It is a global optimization problem.

A MOLFP problem can be formulated as follows [5]:

$$\max\left\{z_1 = \frac{c^1 x + \alpha_1}{d^1 x + \beta_1}\right\} \quad \cdots \quad \max\left\{z_p = \frac{c^p x + \alpha_p}{d^p x + \beta_p}\right\}$$
$$s.t.\ x \in S = \{x \in \mathrm{R}^n | Ax = b, x \geq 0\} \tag{1}$$

where $c^k, d^k \in \mathrm{R}^n$, $\alpha_k, \beta_k \in \mathrm{R}$, $k = 1, \ldots, p$; $A \in \mathrm{R}^{m \times n}$; $b \in \mathrm{R}^m$. We assume that $\forall_{x \in S}, k = 1, \ldots, p : d^k x + \beta_k > 0$ and that S is bounded.

A point $x^1 \in S$ is *weakly-nondominated* if and only if there does not exist another point $x \in S$ such that $z_k(x) > z_k\left(x^1\right)$ for all $k = 1, \ldots, p$. A point $x^1 \in S$ is *nondominated* if and only if there does not exist another point $x \in S$ such that $z_k(x) \geq z_k\left(x^1\right)$, for all $k = 1, \ldots, p$ and $z_k(x) > z_k\left(x^1\right)$ for at least one k. It must be noted that nondominated solutions are also weakly-nondominated but the other weakly-nondominated solutions are dominated.

The *ideal point*, $\tilde{z} \in \mathrm{R}^p$, is the point of the objective function space whose coordinates are equal to the maximum that can be achieved by each objective function in the feasible region, S. $\tilde{z}$ is determined by computing the pay-off table, that is $\tilde{z}^k = z\left(\tilde{x}^k\right)$, $k = 1, \ldots, p$, where $\tilde{x}^k$ optimizes the program:

$$\max\ z_k(x),\ s.t.\ x \in S \tag{2}$$

The point $\tilde{z}$ is found along the main diagonal of the pay-off table, whose rows correspond to the values of the vectors $z\left(\tilde{x}^k\right)$, $k = 1, \ldots, p$.

The general purpose of dealing with multiobjective problems is to compute nondominated solutions that fulfill the decision maker's preferences. In this paper we use *reference point* techniques to provide the framework to specify those preferences. The underlying principle is that a reference point is a goal, but as stated by [6], the meaning of "coming close" changes to "coming close or better", which does not mean the minimization of a distance, but the optimization of an ASF. Such function should ensure non-dominance of the outcome even when the goals are feasible. We use the following ASF to be minimized:

$$s(\bar{z}, z) = \max_{k=1,\ldots,p} (\bar{z}_k - z_k) - \varepsilon \sum_{k=1}^{p} z_k \tag{3}$$

The point $\bar{z} \in \mathrm{R}^p$ is a reference point in the objective function space and it is defined according to the decision maker's preferences. The term $\varepsilon \sum_{k=1}^{p} z_k$ is used to guarantee that the result is a nondominated (and not only weakly-nondominated) solution; ε is a positive number so small that it enables to determine any interesting nondominated solution.

The minimization of the ASF yields a nondominated solution (supported or unsupported) of the MOLFP problem. The ASF problem is:

$$\min\left\{\max_{k=1,\ldots,p}\left[\bar{z}_k - z_k(x)\right] - \varepsilon\sum_{k=1}^{p} z_k(x)\right\},\ s.t.\ x\in S \tag{4}$$

[4] shows that this objective function with $\varepsilon = 0$ is strictly quasi-convex, which means that the problem can be solved with common non-linear techniques, because the local minima are also global minima. However, with $\varepsilon = 0$ the outcome of the multiobjective program can be dominated (weakly nondominated). With $\varepsilon > 0$ the result is a nondominated solution, but the sum of the objective functions (linear fractional) leads to a very hard problem to solve.

We devised a Branch & Bound technique [2] to solve the ASF problem with $\varepsilon > 0$. This technique is based on the fact that the optimization of just one objective function (one fractional ratio) is a convex problem that can be solved, after some variable change techniques [1], by a linear programming algorithm (such as the simplex method). The pay-off table of the MOLFP problem is computed and then the nondominated region is divided into sub-regions by using linear constraints. These sub-regions are then characterized by the computation of their pay-off tables. A region can be discarded if its ideal point is worse than an already computed solution; otherwise it is further sub-divided. The process ends when the ranges of values in the pay-off tables of all non-discarded sub-regions are lower than a certain *error* (tolerance).

In this paper we present a new technique, a Branch & Cut, where special cuts are introduced into the Branch & Bound technique. The cuts stem from some conditions that identify parts of the feasible region where the nondominated solution that optimizes the ASF cannot be obtained. Sect. 3 presents these conditions. Sect. 2 outlines the new technique. The cuts substantially improve the performance of the previous algorithm as it is showed in Sect. 4, where we present the results of several tests.

2 The Cut Conditions

Theorem 1 *Let* $z^1 = z\left(x^1\right)$, $x^1 \in S$, *be a known nondominated solution of the MOLFP problem and* z^+ *a nondominated solution that minimizes* $s(\bar{z},z)$ *over* S. *Consider a region* $T \subseteq S$ *and let* $\tilde{z}^T$ *be its ideal point. Define:*

$$\widehat{z}_r = \bar{z}_r - \max_{k=1,\ldots,p}\left(\bar{z}_k - z_k^1\right) + \varepsilon\sum_{k=1}^{p}\left(z_k^1 - \tilde{z}_k^T\right),\ r = 1,\ldots,p \tag{5}$$

$$C = T \cap \left\{x : z_r(x) \leq \widehat{z}_r,\ r = 1,\ldots,p\right\} \tag{6}$$

Then, either z^1 *minimizes* $s(\bar{z},z)$ *over* S *or* $z^+ \notin C$.

Proof. Let $\tilde{z}^C$ be the ideal point of region C.

For all $x \in C$:

$$z_r(x) \leq \tilde{z}_r^C \leq \widehat{z}_r = \bar{z}_r - \max_{k=1,\ldots,p} \left(\bar{z}_k - z_k^1\right) + \varepsilon \sum_{k=1}^{p} \left(z_k^1 - \tilde{z}_k^T\right),\ r = 1,\ldots,p$$

$$\max_{k=1,\ldots,p} \left(\bar{z}_k - z_k^1\right) - \varepsilon \sum_{k=1}^{p} z_k^1 \leq \left(\bar{z}_r - \tilde{z}_r^C\right) - \varepsilon \sum_{k=1}^{p} \tilde{z}_k^T,\ r = 1,\ldots,p$$

$$s\left(\bar{z}, z^1\right) \leq \left(\bar{z}_r - \tilde{z}_r^C\right) - \varepsilon \sum_{k=1}^{p} \tilde{z}_k^T,\ r = 1,\ldots,p$$

As $s(\bar{z}, z^+) \leq s\left(\bar{z}, z^1\right)$, because z^+ is a solution that minimizes $s(\bar{z}, z)$, then $s(\bar{z}, z^+) \leq \left(\bar{z}_r - \tilde{z}_r^C\right) - \varepsilon \sum_{k=1}^{p} \tilde{z}_k^T$. Moreover, $\tilde{z}_r^T \geq \tilde{z}_r^C, r = 1,\ldots,p$, because $C \subseteq T$, and $\left(\bar{z}_r - \tilde{z}_r^C\right) \leq \max_{k=1,\ldots,p} \left(\bar{z}_k - \tilde{z}_k^C\right)$. Therefore:

$$s\left(\bar{z}, z^+\right) \leq s\left(\bar{z}, z^1\right) \leq \max_{k=1,\ldots,p} \left(\bar{z}_k - \tilde{z}_k^C\right) - \varepsilon \sum_{k=1}^{p} \tilde{z}_k^C = s\left(\bar{z}, \tilde{z}^C\right)$$

Hence, either $s(\bar{z}, z^+) = s\left(\bar{z}, z^1\right)$ or $s(\bar{z}, z^+) < s\left(\bar{z}, \tilde{z}^C\right)$. In the first case, z^1 minimizes $s(\bar{z}, z(x))$. In the second case, $s(\bar{z}, z^+) < s(\bar{z}, z(x)),\ \forall_{x \in C}$, because $\tilde{z}^C$ is the ideal point of region C, thus $z^+ \notin C$. $\square$

3 Outline of the Branch & Cut Algorithm

Initialization. Characterize the nondominated region of the MOLFP problem by computing its pay-off table. Choose the best solution (to be the incumbent) according to the ASF among the solutions of the pay-off table.

Branch & Cut Process. i) Select a region T (in the first iteration, $T = S$) and introduce cuts for all objective functions, which are formulated according to the conditions of Sect. 3: introduce in T the constraints $z_r(x) \geq \widehat{z}_r,\ r = 1,\ldots,p$, with $\widehat{z}_r$ given by (11), where z^1 is the current incumbent solution. ii) Divide this region into two sub-regions, by imposing constraints on one of the objective functions; this objective is the one that has the biggest range of values in the current region, considering both the introduced cuts and the values of the pay-off table; one constraint defines a sub-region where the values of that objective function are greater than the middle value of the range; the other constraint defines another sub-region where the values of that objective function are smaller than the middle value of the range. iii) Compute the pay-off tables of the two new sub-regions in order to characterize them. iv) Update the incumbent solution if any of the new computed solutions is better (according to the ASF) than the previous incumbent solution.

Discarding Condition. A sub-region can be discarded if the value of the ASF for its ideal point is worse than the value of the ASF for the incumbent solution. The demonstration of this condition can be found in [2].

Stopping Condition. The Branch & Cut process is repeated for every non-discarded sub-region until the remaining sub-regions are 'smaller' than a predefined error. One sub-region is 'smaller' than a predefined error when the range of values of each objective function in the pay-off table is lower than the predefined error.

4 Computational Tests

Both algorithms (the Branch & Bound and the Branch & Cut) were implemented with Delphi Pascal 2007 for Microsoft Windows. A laptop T9500, 2.6GHz, with 4 GB RAM, was used for running the tests. The code for solving the linear problems needed by the algorithms was the *lp_solve 5.5*.

The tested problems were randomly generated according to [3]: $c_j^k, d_j^k \in [0.0, 0.5]$ and $a_{ij} \in [0.0, 1.0]$ are uniformly distributed numbers and b equals to 1 . In [3] all constant terms of denominators and numerators are the same number, which ranges from 2.0 and 100.0. Instead, in our tests $\alpha_k, \beta_k \in [2.0, 100.0]$ are also uniformly generated numbers. All the reported tests were carried out with the reference point coordinates set to zero.

Both the Branch & Bound (B & B) and the Branch & Cut (B & C) algorithms were applied to the same problems generated at random: 20 instances for each number of objective functions (from 2 to 10). Each reported measure value is the average of the values obtained through the twenty instances. We used two performance measures: the running time (in seconds) and the number of generated sub-regions. The number of sub-regions is a measure that gives a better indication of the growing complexity of the computations with the different parameters.

Table 1 Performance of the algorithms varying the number of objective functions.

No Object. Funct.	2	3	4	5	6	7	8	9	10
Time B & B	0.2	0.6	2.1	4.7	10.0	18.3	28.0	63.0	197.2
Time B & C	0.1	0.2	0.4	0.5	0.8	1.1	1.4	2.1	4.4
Time Improv.	50%	67%	81%	89%	92%	94%	95%	97%	98%
No Regions B & B	40	82	211	364	646	960	1301	2534	6753
No Regions B & C	14	20	38	43	57	63	76	101	168
No Regions Improv.	65%	76%	82%	88%	91%	93%	94%	96%	98%

Table 1 presents the running time in seconds for both algorithms, considering the number of objective functions growing from two to ten. It also presents the corresponding number of sub-regions generated by the algorithms. These problems have 50 variables and 50 constraints. The error was set to 0.001. The 'improvement' was calculate in the following way: (B&B measure – B&C measure)/ B&B measure.

Table 2 presents the running time in seconds for both algorithms, considering the error ranging from 10^{-1} to 10^{-5}. It also presents the corresponding number of subregions. The problems have 50 variables, 50 constraints and 5 objective functions. The 'improvement' was calculated in the same way as in Table 1.

Table 2 Performance of the algorithms varying the error.

Error	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Time B & B	0.2	1.1	4.7	13.4	31.8
Time B & C	0.1	0.2	0.5	1.3	2.3
Time Improv.	50%	82%	89%	90%	93%
No Regions B & B	8	78	364	1053	2451
No Regions B & C	6	16	43	118	207
No Regions Improv.	25%	79%	88%	89%	92%

5 Conclusion

We presented a new Branch & Cut technique for computing nondominated solutions in MOLFP by using reference point scalarizing programs. Compared with the previous Branch & Bound algorithm that we have developed, the cuts substantially improve the performance of the algorithm. This improvement makes possible to deal with larger problems in a reasonable computational time and with a common machine.

References

1. A Charnes and W Cooper. Programming with linear fractional functions. *Naval Research Logistics Quarterly*, 9: 181–186, 1962.
2. João Paulo Costa and Maria João Alves. A reference point technique to compute nondominated solutions in MOLFP. *Journal of Mathematical Sciences*, 161(6): 820–831, 2009.
3. T Kuno. A branch-and-bound algorithm for maximizing the sum of several linear ratios. *Journal of Global Optimization*, 22: 155–174, 2001.
4. B Metev and D Gueorguieva. A simple method for obtaining weakly efficient points in multiobjective linear fractional programming problems. *European Journal of Operational Research*, 126: 386–390, 2000.
5. Ralph Steuer. *Multiple Criteria Optimization: Theory, Computation and Application*. Wiley, New York, 1986.
6. Adrezj Wierzbicki. Reference point methods in vector optimization and decision support. Technical Report IR-98-017, IIASA, Laxenburg, 1998.

Using a Genetic Algorithm to Solve a Bi-Objective WWTP Process Optimization

Lino Costa, Isabel A. C. P. Espírito Santo, and Edite M. G. P. Fernandes

Abstract When modeling an activated sludge system of a wastewater treatment plant (WWTP), several conflicting objectives may arise. The proposed formulation is a highly constrained bi-objective problem where the minimization of the investment and operation costs and the maximization of the quality of the effluent are simultaneously optimized. These two conflicting objectives give rise to a set of Pareto optimal solutions, reflecting different compromises between the objectives. Population based algorithms are particularly suitable to tackle multi-objective problems since they can, in principle, find multiple widely different approximations to the Pareto-optimal solutions in a single run. In this work, the formulated problem is solved through an elitist multi-objective genetic algorithm coupled with a constrained tournament technique. Several trade-offs between objectives are obtained through the optimization process. The direct visualization of the trade-offs through a Pareto curve assists the decision maker in the selection of crucial design and operation variables. The experimental results are promising, with physical meaning and highlight the advantages of using a multi-objective approach.

1 Multi-Objective Optimization

We apply the Multi-objective Elitist Genetic Algorithm (MEGA), described in [3] to the WWTP multi-objective optimization problem. This approach, in contrast to other algorithms, does not require any differentiability or convexity conditions of the search space. Moreover, since it works with a population of points, it can find, in a single run, multiple approximations to the solutions of the Pareto optimal set without the need of fixing any weights and a well distributed representation of the Pareto optimal frontier induced by the use of diversity-preservation mechanisms.

Lino Costa · Isabel A. C. P. Espírito-Santo · Edite M. G. P. Fernandes
Department of Production and Systems, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal e-mail: {lac,iapinho,emgpf}@dps.uminho.pt

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_57,

We now shortly describe some technical features and the parameters of the MEGA algorithm (Algorithm 1). MEGA starts from a population of points P of size s. In

Algorithm 2 Multi-objective Elitist Genetic Algorithm

Require: $e \geq 1$, $s > 1$, $0 < p_c < 1$, $0 < p_m < 1$, $s_{SP} > s$, $\sigma_{\text{share}} > 0$
1: $k \leftarrow 0$
2: **for** $l = 1, \ldots, s$ **do**
3: Randomly generate the main population P
4: **end for**{(Initialization of the population)}
5: **while** stopping criterion is not met **do**
6: Fitness assignment $FA(P, \sigma_{\text{share}})$ for all points in main population P
7: Update the secondary population SP with the non-dominated points in P
8: Introduce in P the elite with e points selected at random from SP
9: Select by tournaments s points from P
10: Apply SBX crossover to the s points, with probability p_c
11: Apply mutation to the s points with probability p_m
12: $k \leftarrow k + 1$
13: **end while**
14: Update the secondary population SP with the non-dominated points in P
15: **return** Non-dominated points from SP

our implementation, a real representation is used since we are leading with a continuous problem. Additionally, a secondary population SP that archives potential Pareto optimal solutions found so far during the search process is maintained. The elitist technique implemented is based on the secondary population with a fixed parameter e ($e \geq 1$) that controls the elitism level, i.e., e is the maximum number of non-dominated solutions of the secondary population that will be introduced in the main population. These non-dominated solutions will effectively participate in the search process that is performed using the points of the main population.

In order to handle constraints, we implemented the constrained tournament method in which a new dominance relation is defined [4]. A solution $x \in \mathbb{R}^n$ constrain-dominates $y \in \mathbb{R}^n$, i.e., $x \prec_c y$ if and only if: x is feasible and y is not; x and y are unfeasible, but x has a smaller constraint violation; x and y are feasible, x dominates y, i.e., $x \prec y$.

Solutions are evaluated according to a fitness assignment function $FA(P, \sigma_{\text{share}})$ that is based on the constraint-dominance relation between points. All solutions are ranked in terms of dominance defining several fronts. Therefore, all non-dominated solutions in the main population P will constitute the first front to which is assigned a rank equal to 1. Successively, the same procedure is applied to the remaining points defining several fronts with increasing ranks. In order to maintain diversity, a sharing scheme depending on an initial parameter σ_{share} is applied to the solutions belonging to the same front. For this purpose, an adaptive sharing scheme on objective space was adopted for diversity preservation as described in [2].

Non-dominated points in main are archived in SP. The SP update implies the determination of Pareto optimality of all solutions stored so far, in order to eliminate those that became dominated. As the size of SP grows, the time to complete this

operation may become significant. So, in order to prevent the growing computation times, in general, a maximum $s_{SP} > s$ size is imposed.

A tournament selection that guarantees that better points are more likely to be selected was used to select points from the main population. New points in the search space are generated by the application, with probability p_c, of a Simulated Binary Crossover (SBX) [3, 2] that simulates the working principle of single-point crossover operator for binary strings. A Polynomial Mutation is applied, with a probability p_m, to the points produced by the crossover operator. Mutation introduces diversity in the population since crossover, exclusively, could not assure the exploration of new regions of the search space.

The search ends when a given stopping criterion is satisfied. The best approximations to the Pareto-optimal set are archived in SP.

2 The Case Study: WWTP Optimal Design

The system under study consists of an aeration tank, where the biological reactions take place, and a secondary settler for the sedimentation of the sludge and clarification of the effluent. To describe the aeration tank we chose the activated sludge model n.1, described by Henze et al. [8]. The tank is considered to operate in steady state and as a completely stirred tank reactor and the generic equation for a mass balance in these conditions is

$$\frac{Q}{V_{\mathrm{a}}}(\xi_{\mathrm{in}} - \xi) + r_{\xi} = 0,$$

where Q is the flow that enters the tank, V_{a} is the aeration tank volume, ξ and ξ_{in} are the concentrations of the components, particulates or solubles, around which the mass balances are being made inside the reactor and on entry, respectively. r_{ξ} is obtained by the Peterson Matrix [8] and is the process reaction rate.

Another set of constraints is concerned with the secondary settler. When the wastewater leaves the aeration tank, the treated water should be separated from the biological sludge, otherwise, the chemical oxygen demand would be higher than it is at the entry of the system. The ATV design procedure [5] contemplates the peak wet weather flow events, during which there is a reduction in the sludge concentration and is based on very simple empirical relations.

Besides the ATV procedure, the double exponential model [9] is also used to describe the sedimentation process [7]. This model assumes a one dimensional settler, in which the tank is divided into 10 layers of equal thickness. It assumes that no biological reactions take place, meaning that the dissolved matter concentration is maintained across all the layers. Only vertical flux is considered and the solids are uniformly distributed across the entire cross-sectional area of the feed layer. This model is based on a traditional solids flux analysis but the flux in a particular layer is limited by what can be handled by the adjacent layer. The settling function is given by

$$v_{s,j} = \max\left(0, \min\left(v_0', v_0\left(e^{-r_h(TSS_j - f_{ns}TSS_a)} - e^{-r_p(TSS_j - f_{ns}TSS_a)}\right)\right)\right)$$

where $v_{s,j}$ is the settling velocity in layer j, TSS_j is the total suspended solids concentration in each of the ten considered layers and v_0, v_0', r_h, r_p and f_{ns} are settling parameters [7]. This model introduces discontinuities in the problem.

The other important group of constraints are a set of linear equalities and defines composite variables. In a real system, some state variables are, most of the time, not available for evaluation. Thus, readily measured composite variables are used instead. This includes the chemical oxygen demand (COD), total suspend solids (TSS) and total nitrogen (N), to name the more important.

The system behavior, in terms of concentration and flows, may be predicted by balances. In order to achieve a consistent system, these balances must be done around the entire system and not only around each unitary process. They were done to the suspended matter, dissolved matter and flows. For example, to the soluble compounds, represented by $S_?$ we have

$$(1+r)Q_{\text{inf}}S_{?\text{ent}} = Q_{\text{inf}}S_{?\text{inf}} + rQ_{\text{inf}}S_?$$

where r is the recycle rate and $Q_?$ the volumetric flows. As to the subscripts, " inf " concerns the influent wastewater and " ent " the entry of the aeration tank.

It is also necessary to add some system variables definitions, in order to define the system correctly. All the variables are considered non-negative, although more restricted bounds are imposed to some of them due to operational consistencies. As an example, the amount of soluble oxygen in the aeration tank must be at least 2 g/mL. These conditions define a set of simple bounds on the variables.

Finally, the quality of the effluent has to be imposed. The quality constraints are usually derived from law restrictions. The most used are related with limits in the Chemical Oxygen Demand, Nitrogen and Total Suspended Solids at the effluent. In mathematical terms, these constraints are defined by Portuguese laws as $COD_{\text{ef}} \leq 125$, $N_{\text{ef}} \leq 15$ and $TSS_{\text{ef}} \leq 35$. We refer to [6] for more details.

The first objective function of the problem represents the total cost and includes both investment and operation costs. The operation cost is usually on annual basis, so it has to be updated to a present value using the adequate economic factors of conversion. Each term in the objective function is based on the basic model $C = aZ^b$ [10], where a and b are the parameters to be estimated, C is the cost and Z is the characteristic of the unitary process that most influences the cost. The parameters a and b are estimated by the least squares technique, using real data collected from a WWTP building company. Summing up the terms from all the costs in all considered units, we get the following Total Cost (TC) objective function that depends on V_{a}, the air flow (G_{S}), the sedimentation area (A_{s}) and depth (h).

$$TC(V_{\text{a}}, G_{\text{S}}, A_{\text{s}}, h) = 174.2V_{\text{a}}^{1.07} + 12487G_{\text{S}}^{0.62} + 114.8G_{\text{S}} + 955.5A_{\text{s}}^{0.97} + 41.3\,(A_{\text{s}}h)^{1.07} \quad .$$

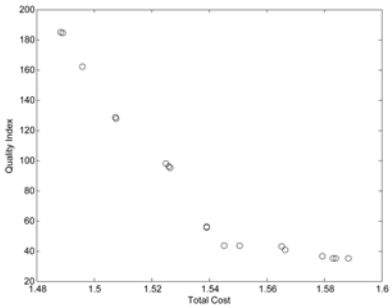
The second objective function is the Quality Index (QI) defined by the BSM1 model [1] and gives a measure of the amount of daily pollution.

It depends on the quality of the effluent in terms of TSS, COD, biochemical oxygen demand (BOD), total Kjeldahl nitrogen (TKN), nitrate and nitrite nitrogen (S_{NO}) and the effluent flow (Q_{ef}). The obtained function is

$$QI(TSS,COD,BOD,TKN,NO,Q_{ef}) = (2TSS+COD+2BOD+20TKN+2S_{NO})\frac{Q_{ef}}{1000}.$$

3 Numerical Results and Conclusions

The mathematical model has 2 objective functions, 71 parameters, 113 variables, 103 equality constraints and one inequality constraint. All the variables are bounded below and above. The stoichiometric, kinetic and operational parameters are the default values presented in the GPS-X simulator [11], and they are usually found in real activated sludge based plants. The MatLab implementation of the problem is available from the webpage http://www.norg.uminho.pt/iapinho/proj.htm under "Bi-objective WWTP Project".



	TC_{min}	QI_{min}
V_a	1567	1567
G_S	100	100
A_s	816	813
h	3.3	4.9
TC	1.49	1.59
QI	185	35

Fig. 1 Pareto curve for the Total Cost and Quality Index, and optimal values for the most important variables

The MEGA algorithm was coded in MatLab programming language and the numerical results were obtained with a Intel Core2 Duo CPU 1.8GHz with 2GB of memory. The MEGA parameters are: $s = 40$, $e = 4$, $p_c = 0.9$, $p_m = 1/113$, $s_{SP} = \infty$ and $\sigma_{share} = 0.1$. The maximum number of objective function evaluations is 50000. An initial value, x^0, provided by the GPS-X simulator [11] with the real influent data was introduced in the initial population. Several experiments were conducted without introducing this initial point in the population and the algorithm failed to achieve a feasible point within the maximum number of objective function evaluations.

Figure 1 shows the Pareto optimal front defined by the approximations to the Pareto optimal solutions. In this figure, the compromise solutions between QI and TC are plotted. It is also presented the results for the most important decision variables of the limit solutions from the Pareto front (TC_{min} and QI_{min}), namely, the

aeration tank volume, the air flow, the area and depth of the secondary settler, as well as TC and QI. The total computational time was about 190 seconds. We can observe that the non-dominated solutions obtained are viable and have physical meaning, highlighting the superiority of the bi-objective approach in terms of computational demands. The obtained WWTP designs represent compromises that are economically attractive with convenient quality indexes and satisfy the law limits. Moreover, these limits in terms of COD and TSS are below the law limits ($COD_{TC_{\min}} = 36.1$, $COD_{QI_{\min}} = 22.6$, $TSS_{TC_{\min}} = 12.2$ and $TSS_{QI_{\min}} = 6.6$), showing the robustness of the solution. Although the obtained WWTP designs are attractive, in the future we intend to propose a multi-objective approach with more than two objectives. For example, air flow requirements and chemicals addition will be considered.

References

1. J. Alex, L. Benedetti, J. Copp, K.V. Gernaey, U. Jeppsson, I. Nopens, M.N. Pons, C. Rosen, J.P. Steyer, and P. Vanrolleghem (2008) Benchmark Simulation Model no. 1 (BSM1). *Technical Report prepared by the IWA Taskgroup pn Benchmarking of Control Strategies for WWTPs*.
2. L. Costa and P. Oliveira (2003) An Adaptive Sharing Elitist Evolution Strategy for Multiobjective Optimization. *Evolutionary Computation*, 11(4), 417–438.
3. L. Costa and P. Oliveira (2003) An elitist genetic algorithm for multiobjective optimization, in M.G.C. Resende and J.P. de Sousa (eds.), *Metaheuristics: Computer Decision-Making*, pp. 217–236, Kluwer Academic Publishers.
4. K. Deb (2000) An efficient constraint handling method for genetic algorithms, *Computer Methods in Applied Mechanics and Engineering*, 186(2–4), 311–338.
5. G.A. Ekama, J.L. Barnard, F.H. Günthert, P. Krebs, J.A. McCorquodale, D.S. Parker, and E.J. Wahlberg (1978) *Secondary Settling Tanks: Theory, Modeling, Design and Operation*. Technical Report 6, IAWQ - International Association on Water Quality.
6. I.A.C.P. Espírito Santo, E.M.G.P. Fernandes, M.M. Araújo, and E.C. Ferreira (2006) How Wastewater Processes can be Optimized using LOQO. *Lecture Notes in Economics and Mathematical Systems*, A. Seeger (ed.), Springer-Verlag 563: 435–455.
7. I.A.C.P. Espírito Santo, E.M.G.P. Fernandes, M.M. Araújo, and E.C. Ferreira (2006) On the Secondary Settler Models Robustness by Simulation. *WSEAS Transactions on Information Science and Applications* 3: 2323–2330.
8. M. Henze, C.P.L Grady Jr, G.V.R Marais, and T. Matsuo (1986) *Activated Sludge Model no. 1*, Technical Report 1, IAWPRC Task Group on Mathematical Modelling for Design and Operation of Biological Wastewater Treatment.
9. I. Takács, G.G. Patry, and D. Nolasco (1991) A Dynamic Model of the Clarification-Thickening Process, *Water Research* 25: 1263–1271.
10. D. Tyteca, Y. Smeers, and E.J. Nyns (1977) Mathematical Modeling and Economic Optimization of Wastewater Treatment Plants. *CRC Critical Reviews in Environmental Control* 8: 1–89.
11. GPS-X Technical Reference (2002). Hydromantis, Inc.

The q-Gradient Vector for Unconstrained Continuous Optimization Problems

Aline Cristina Soterroni, Roberto Luiz Galski, and Fernando Manuel Ramos

Abstract In the beginning of nineteenth century, Frank Hilton Jackson generalized the concepts of derivative in the q-calculus context and created the q-derivative, widely known as Jackson's derivative. In the q-derivative, the independent variable is multiplied by a parameter q and in the limit, $q \to 1$, the q-derivative is reduced to the classical derivative. In this work we make use of the first-order partial q-derivatives of a function of n variables to define here the q-gradient vector and take the negative direction as a new search direction for optimization methods. Therefore, we present a q-version of the classical steepest descent method called the q-steepest descent method, that is reduced to the classical version whenever the parameter q is equal to 1. We applied the classical steepest descent method and the q-steepest descent method to an unimodal and a multimodal test function. The results show the great performance of the q-steepest descent method, and for the multimodal function it was able to escape from many local minima and reach the global minimum.

1 Introduction

It is well-known that along the direction given by the gradient vector the objective function $f(x)$ increases most rapidly. If the optimization problem is to minimize an objective function, then it is intuitive to use the steepest descent direction $-\nabla f(x)$ as the search direction in optimization methods. Here we introduce the q-gradient vector that is similar to the classical gradient vector, but instead of the classical first-order partial derivatives we use the first-order partial q-derivatives obtained from the Jackson's derivative, also referred to as q-difference operator, or q-derivative operator or simply q-derivative.

Aline Cristina Soterroni - Roberto Luiz Galski - Fernando Manuel Ramos
Applied Computing Program - Satellite Control Centre - Computing and Applied Mathematics Associated Laboratory, National Institute for Space Research, Av. dos Astronautas, 1758, 12227-010, Brazil. e-mail: aline.soterroni@lac.inpe.br,galski@ccs.inpe.br,fernando@lac.inpe.br

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_58,

Frank Hilton Jackson gave many contributions related to basic analogues or q-analogues, especially on basic hypergeometric functions [1], and in the beginning of nineteenth century he generalized the concepts of derivative and integration in the q-calculus context and created the q-derivative and the q-integration [3, 4]. In the q-derivative, instead of the independent variable x of a function $f(x)$ be added by a infinitesimal value, it is multiplied by a parameter q that is a real number different from 1. And in the limit, $q \to 1$, the q-derivative tends to the classical derivative. Our proposal approach has two advantages. On the one hand, the closer to 1 the value of the parameter q is, the closer to the classical gradient vector the q-gradient vector will be. For monomodal poorly scaled functions, to use a direction close, but not equal, to the steepest descent direction can reduce the zigzag movement towards the solution. On the other hand, when $q \neq 1$ the q-gradient vector can make any angle with the classical gradient vector and the search direction can point to any direction. It can be interesting for multimodal functions because the search procedure can escape from the many local minima in the search space. The main objective of this work is to introduce the q-gradient vector based on the q-derivative and use its negative as search direction in optimization methods for unconstrained continuous optimization problems. The paper is organized as follows. In Section 2 the q-gradient vector is defined. In Section 3 a q-version of the classical steepest descent method is presented. Section 4 deals with the performance of the q-steepest descent method and the classical steepest descent method for numerical examples. And Section 5 contains final considerations.

2 q-Gradient Vector

Let $f(x)$ be a real-valued continuous function of a single variable, the q-derivative of f is given by

$$D_q f(x) = \frac{f(x) - f(qx)}{(1-q)x}, \tag{1}$$

with $x \neq 0$ and $q \neq 1$. And in the limiting case, $q = 1$, the q-derivative is equal to the classical derivative provided that f is differentiable at x. The parameter q is usually taken as a fixed real number $0 < q < 1$, but the q can be a real number different from 1. Note, in the Equation (1), the q-derivative is not defined at $x = 0$. Therefore, for real-valued continuous functions differentiable at $x = 0$, the q-derivative can be given by [5]

$$D_q f(x) = \begin{cases} \frac{f(x)-f(qx)}{(1-q)x}, & x \neq 0, \ q \neq 1 \\ \frac{df(0)}{dx}, & x = 0 . \end{cases} \tag{2}$$

For a real-valued continuous function of n variables, $f(\mathbf{x})$, the gradient vector is the vector of the n first-order partial derivatives of f provided that the function has first-order partial derivatives with respect to all independent variables x_i ($i =$

$1,2,\cdots,n$). Similarly, the q-gradient vector is the vector of the n first-order partial q-derivatives of f. Before introducing the q-gradient vector it is convenient to define the first-order partial q-derivative of a real-valued continuous function of n variables with respect to a variable x_i as [6]

$$D_{q_i,x_i}f(\mathbf{x}) = \frac{f(x_1,\ldots,x_{i-1},x_i,x_{i+1}\ldots,x_n) - f(x_1,\ldots,x_{i-1},q_ix_i,x_{i+1},\ldots,x_n)}{(1-q_i)x_i}, \quad (3)$$

with $x_i \neq 0$ e $q_i \neq 1$. Similarly to the Equation (2), we can define the first-order partial q-derivative with respect to a variable x_i as follow

$$D_{q_i,x_i}f(\mathbf{x}) = \begin{cases} \frac{f(x_1,\ldots,x_{i-1},x_i,x_{i+1}\ldots,x_n)-f(x_1,\ldots,x_{i-1},q_ix_i,x_{i+1}\ldots,x_n)}{(1-q_i)x_i}, & x_i \neq 0,\ q \neq 1 \\ \frac{\partial f}{\partial x_i}(x_1,\ldots,x_{i-1},0,x_{i+1}\ldots,x_n), & x_i = 0 \\ \frac{\partial f}{\partial x_i}(x_1,\ldots,x_{i-1},x_i,x_{i+1}\ldots,x_n), & q_i = 1\,. \end{cases} \quad (4)$$

Therefore, considering the parameter $\mathbf{q} = (q_1,\ldots,q_i,\ldots,q_n)$ and let $f(\mathbf{x})$ be a real-valued continuous function which has first-order partial q-derivatives with respect to all independent variables x_i ($i = 1,2,\cdots,n$), we introduce here the q-gradient vector as

$$\nabla_{\mathbf{q}} f(\mathbf{x}) = \left[D_{q_1,x_1}f(\mathbf{x}) \ \ldots\ D_{q_i,x_i}f(\mathbf{x}) \ \ldots\ D_{q_n,x_n}f(\mathbf{x})\right], \quad (5)$$

where $D_{q_i,x_i}f(\mathbf{x})$ is given by the Equation (4). And in the limit, $q_i \to 1$, for all i ($i = 1,\ldots,n$), the q-gradient vector returns to the usual gradient vector.

3 q-Steepest Descent Method

A general optimization strategy is to consider an initial set of independent variables, $\mathbf{x}_0$, and apply an iterative procedure given by $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_k\mathbf{s}_k$, where k is the iteration number, $\mathbf{x}$ is the vector of variables, α is the step length and $\mathbf{s}$ is a search direction vector. This process continues until either no additional reduction in the value of the objective function can be made or the solution point has been approximated with sufficient accuracy [8]. In the steepest descent method the search direction $\mathbf{s}_k$ is given by the negative of the gradient vector at the point $\mathbf{x}_k$, $-\nabla f(\mathbf{x_k})$, and the step length α_k can be found by a one-dimensional search performed in the direction $\mathbf{s}_k$. Similarly, the search direction for the q-steepest descent method is given by the negative of the q-gradient vector at the point $\mathbf{x}_k$, $-\nabla_{\mathbf{q}} f(\mathbf{x_k})$ (see Equation (5)).

A possible strategy to compute the parameters q_i ($i = 1,2,\ldots,n$) is to generate random numbers from a log-normal distribution with a fixed mean $\mu = 1$ and a variable standard deviation σ. The log-normal distribution has multiplicative effects and the q-gradient vector is related to multiplications instead of additions in the independent variables. When the multiplication is the relevant operation for

combining quantities, the log-normal distribution is more indicated [2]. In our strategy, the initial standard deviation σ_0 should be a real positive number different from zero and during the iterative procedure it is reduced to zero by $\sigma_k = \beta \cdot \sigma_{k-1}$, where $0 < \beta < 1$. In the beginning of the iterative procedure, when $\sigma_k \neq 0$ with $\mu = 1$, this strategy implies the parameters q_i can be any real positive number with more occurrence around the mean, but with the same likelihood to occur in the intervals $(0,1)$ or $(1,\infty)$. At that time, the q-gradient vector can point to any direction. It gives the method the possibility to search in other directions different from the steepest descent direction and escape from local minima for multimodal functions, or reduce the zigzag movement towards the minimum for poorly scaled functions. At the end of the iterative procedure, when σ_k tends to 0 with $\mu = 1$, the parameters q_i tend to 1 and the q-gradient vector tends to the usual gradient vector. In other words, when $\sigma \to 0$ this strategy makes the q-steepest descent method be reduced to the classical steepest descent method. The optimization algorithm for the q-steepest descent method is given below.

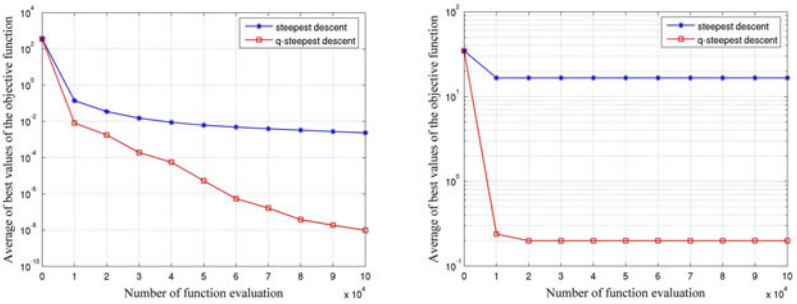
Algorithm 1
Step 1: Initialize randomly $\mathbf{x}_0 \in \mathbb{R}^n$, set $\mu = 1$, and take σ_0 and β.
Step 2: Set $k := 1$.
Step 3: Generate the parameter $\mathbf{q} = (q_1, \ldots, q_i, \ldots, q_n)$ by a log-normal distribution with mean μ and standard deviation σ_k.
Step 4: Compute the search direction $\mathbf{s}_k = -\nabla_q f(\mathbf{x}_k)$.
Step 5: Find the step length α_k by one-dimensional search.
Step 6: Compute $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k$.
Step 7: If the stopping conditions are reached, then stop; Otherwise go to Step 8.
Step 8: Reduce the standard deviation $\sigma_k = \beta \cdot \sigma_{k-1}$.
Step 9: Set $k := k+1$, and go to Step 3.

4 Numerical Examples

We applied the classical steepest descent method and the q-version presented here (Algorithm 1) for the same functions, initial set of independent variables, stopping condition, and strategy to find the step length. We considered the unimodal Rosenbrock function and the highly multimodal Rastrigin function with two variables (see [7] for more details). We generated 50 different initial points $\mathbf{x}_0$ from a uniform distribution in the interval $(-2.048, 2.048)$ for the Rosenbrock function, and the interval $(-5.12, 5.12)$ for the Rastrigin function. The stopping condition was the maximum number of function evaluations equal to 10^5. And for the one-dimension searches we used the golden section method by the code found in [9]. For the numerical derivatives we used the Ridder's method whose code is also found in [9]. Besides that, we considered the parameter q_i numerically equal to 1 when $q_i \in [1-\varepsilon, 1+\varepsilon]$ with ε equal to 10^{-4}.

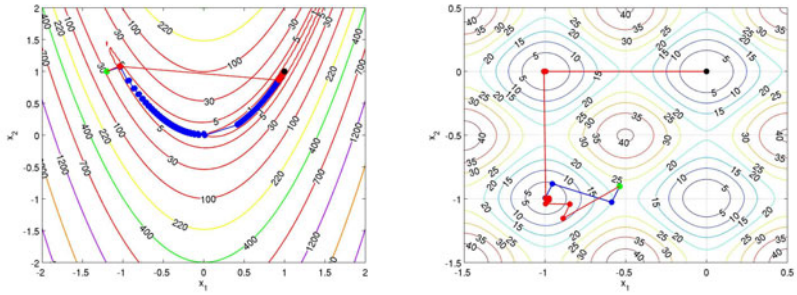
For the Rosenbrock function we used the mean $\mu = 1$ and initial standard deviation $\sigma_0 = 0.5$ with reduction factor $\beta = 0.999$. For the Rastrigin function we

set the mean $\mu = 1$ and initial standard deviation $\sigma_0 = 5.0$ with reduction factor $\beta = 0.995$. The performance results are shown in Fig. 1. The Fig. 2 displays the trajectories of the steepest descent method (blue line) and the q-steepest descent method (red line) for selected initial points.



(a) (b)

Fig. 1 Average of the best function values versus the number of function evaluations for the test functions (a) Rosenbrock and (b) Rastrigin.



(a) (b)

Fig. 2 Trajectories of the steepest descent method (blue line) and the q-steepest descent method (red line) for (a) Rosenbrock and (b) Rastrigin functions. The green point is the initial point and the black point is the global minimum.

It can be seen from the Fig. 1(a) that the steepest descent method over the Rosenbrock function converges more slowly than the q-steepest descent method. The average of the best function values at 10^5 function evaluations is equal to $2.31 \cdot 10^{-3}$ for the steepest descent method and $9.38 \cdot 10^{-9}$ for the q-steepest descent method. For the Rastrigin function, Fig. 1(b), after some function evaluations the curve for the steepest descent method is a straight line because the method moved toward the local minimum closest to each 50 initial points and remained there until the stopping condition has been reached. For the q-steepest descent method, 42 initial points reached the global minimum. The other 8 points that moved toward local minima make the curve for the q-steepest descent method in the Fig. 1(b) has a

small variation after some function evaluations. The average of the best function values at 10^5 function evaluations is equal to 16.5958 for the steepest descent method and for the q-steepest descent method it is equal to 0.1989. The Fig. 2(a) shows that the q-steepest descent method for the Rosenbrock function can reduce the zigzag movement towards the minimum, and for the Rastrigin function, Fig. 2(b), the method can escape from the local minima and reach the global minimum of this multimodal function.

5 Final Considerations

In this work we introduced the q-gradient vector based on the first-order partial q-derivatives of a function of n variables. In addition, we presented a q-version for the steepest descent method with a possible strategy to generate the parameter q. The performance results show the advantage of using the negative of the q-gradient vector as the search direction in optimization methods for unconstrained continuous problems. Particularly, for the multimodal function it could be seen that the q-steepest descent method has mechanisms to escape from the many local minima and move towards the global minimum. These results show that the idea of using the q-gradient vector instead of the classical one points to ways of generalizing other well-known optimization algorithms, such as the conjugate gradient method, with the help of Jackson's q-calculus.

Acknowledgements This work was supported by National Counsel of Technological and Scientific Development (CNPq), Brazil.

References

1. T. W. Chaundy. Frank Hilton Jackson. *J. London Math. Soc.*, 37: 126–128, 1962.
2. W. A. Stahel, E. Limpert, and M. Abbt. Log-normal distributions across the sciences: keys and clues. *BioScience*, 51: 341–352, 2001.
3. F. H. Jackson. On q-functions and a certain difference operator. *Trans. Roy. Soc. Edinburg*, 46: 253–281, 1908.
4. F. H. Jackson. On q-definite integrals. *Quart. J. Pure Appl. Math.*, 41: 193–203, 1910.
5. J. Koekoev and R. Koekoev. A note on the q-derivative operator. *Journal of Mathematical Analysis and Applications*, 176: 627–634, 1993.
6. S. D. Marinkovic, P. M. Rajkovic, and M. S. Stankovic. On q-Newton-Kantorovich method for solving systems of equations. *Applied Mathematics and Computation*, 168: 1432–1448, 2005.
7. H. Pohlheim. GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with Matlab. *http://www.geatbx.com/docu/fcnindex-01.html*, 1994.
8. G. N. Vanderplaats. *Multidiscipline design optimization*. Vanderplaats Research and Development, INC., Monterey, 2007.
9. W. T. Vetterling, W. H. Press, S. A. Teukolsky, and B. P. Flannery. *Numerical recipes in Fortran 90: The art of parallel scientific computing*. Cambridge University Press, New York, 1996.

II.5 Production and Service Management

Chair: Prof. Dr. Marion Steven (Ruhr-University Bochum)

Although developed economies are said to be changing into service economies, their industrial basis still has a considerable significance. Management activities for production of material goods and services show characteristic similarities and differences. Furthermore, in many industries provision of products and services is not separated but integrated in form of product-service systems. Planning tasks comprise setting up, running and improving these systems by appropriate solution methods.

Possible contributions may cover the fields of operations management, project management, forecasting, design of products and services, quality management, process, location and layout strategies, supply chain management, inventory management, material requirements planning and short-term scheduling of product-service systems using methods like optimization techniques, esp. linear and integer programming, simulation, forecasting and analysis methods.

Flow Shop Scheduling at Continuous Casters with Flexible Job Specifications

Matthias Wichmann, Thomas Volling, and Thomas S. Spengler

Abstract We propose a new approach for the sequencing of slabs at continuous casters. Here, typical characteristics such as input supply in batch quantities, continuous casting width adjustment, width ranges for orders as well as setup times due to several process constraints are taken into account. The aim of the optimization approach is to minimize the total cost of production for a given order set. Thus, classic sequencing approaches do not cover all necessary constraints of the process. We present an extended MIP modeling approach for the introduced decision situation of scheduling at a single machine.

1 Introduction

Continuous casting is a production process in the continuous flow shops of the chemical and metal industry. The input of the production process are liquid goods such as molten steel or copper. During the casting process this liquid material is shaped into a continuous solid bar which is divided into smaller pieces for further production processes. In the steel industry, continuous casting is the most important type of primary shaping, which is used to set the width and the thickness of processed material. The casting process is concluded by cutting the steel bar lengthwise into slabs, cuboid blocks of solid steel. Each of the slabs is assigned to a specific customer order. Thus and due to the high value of material, the casting process is order driven [4].

The task of order driven production planning at continuous casters is to generate a schedule for known production orders. Doing so, the characteristics of the produc-

Matthias Wichmann · Thomas Volling · Thomas S. Spengler
Institute of Automotive Management and Industrial Production, Technische Universität Braunschweig, Katharinenstr. 3, D-38106 Braunschweig, e-mail: ma.wichmann | t.volling | t.spengler@tu-bs.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 373
DOI 10.1007/978-3-642-20009-0_59,

tion process have to be taken into account. These are very specific and do not allow for standard scheduling approaches to be applied.

The aim of the paper is to present a scheduling model for the introduced planning situation. The problem characteristics and the resulting impact on constraints and objective function will be discussed in more detail in Section 2. In Section 3 a scheduling model will be developed, which incorporates the given constraints. To validate the model, an illustrative case study is given in Section 4. The contribution is closed with a short outlook.

2 Scheduling at Continuous Casters

In this section the problem characteristics, which require for an extension of classic scheduling approaches, are discussed in more detail. They can be categorized as follows.

First, the task of continuous casting is to provide the following production stages with the slabs to fulfill specified customer orders. Both, the casting velocity measured by cast meters per minute, and the casting thickness are technologically predetermined. Degrees of freedom exist with respect to the casting width as well as the length of the resulting slabs. Both can be set in order to meet customer orders requirements. Each customer order, in terms of a job to be scheduled, is specified by a steel alloy, a casting weight as equivalent to the volume, and a casting width. Since material can be reformed in consecutive production stages, the width is not fixed to a specific value but to a range, determined by a lower and an upper bound. This results into flexible job specifications. Thus, the assumption of a fixed casting width of each job of known approaches like, e.g. [3, 5] is not sufficient and should be enhanced to a more realistic aspect. Given the volume to fulfil the customer order, the casting width is directly linked to the length of the slab and, as such, to the time needed for casting, as determined by the fixed velocity. Accordingly, the higher the casting width the shorter the slab and the higher is the performance of the caster.

Second, the casting width is a technological parameter and can be changed continuously during the casting process. Thus, the casting width at the beginning of a job can be different to the casting width at the end of the same job. However, process constraints apply. Since the change is limited to a maximum rate, there has to be a smooth transition between two consecutive jobs. Therefore, the end width of the preceding job has to match the start width of the following job. Since the processing time of a job depends on the casting width and the assigned casting width of a job depends on the schedule, the processing time as well as any time based job evaluation is not fixed as assumed in general [2].

Third, material to be cast is supplied in charges of a specific alloy with a weight that by far exceeds a single order's weight. Orders of the same alloy have to be pooled to utilize as much material of a charge as possible for order fulfilment. Thus, due to order driven production, steel that is not assigned to orders can be regarded as scrap. To minimize scrap, charges of the same alloy can be strung together. Doing

so, unused material may be pooled to fulfill further jobs. In common approaches, the assignment of jobs to charges and the sequencing of charges are split into two consecutive planning problems [1]. In our approach we integrate both planning tasks into the order scheduling. Doing so, the scheduling approach needs to be extended by bin packing aspects.

Fourth, as in each scheduling approach, machine setups have to be taken into account. Machine setups arise based on several rules. A setup is required if a stepwise width change between two consecutive jobs is necessary. Thus, all remaining material of the preceeding charge is processed as scrap. Afterwards, the caster stops production and is adjusted to the new width. Finally, production begins with a new charge consisting of alloy for the successing job. Another kind of setup is necessary, if the alloy between two consecutive jobs changes while the width remains unchanged. Since mixing of alloys is not allowed, the remaining material of the preceding charge is processed through the caster and needs to be scrapped. Production begins with a new charge consisting of alloy for the successing job but, as a difference, without stopping production.

The objective of job scheduling at continuous casters is to minimize overall production cost. Cost arise from three factors. The first, most common factor regards costs for setups. The second factor are material cost for scrap, that results from batching jobs to charges. The third cost factor is related to machine utilization. If an order is not cast at its maximum casting width, the production time increases resulting in a decreased caster performance. Thus, the difference between maximum and realized casting width can be evaluated with penalty costs. Due to the short planning horizon of scheduling continuous casters, holding costs are not taken into account.

3 Model Formulation

In this section a mathematical formulation for the scheduling problem is developed. We will use the following notation: For each job i of the n jobs to be scheduled, the weight $weight_i$, minimum width wd_i^{min} and maximum width wd_i^{max} are given. The binary data $fit_{i,i'}$ indicates, whether the required alloys of job i and i' are similar. The weight amount of a charge is $CHARGE$. M refers to a big number.

For a schedule, each job i has to be assigned to an order position j of the schedule with the length of n positions. Doing so, the binary decision variable x_{ij} indicates, whether job i is assigned to the jth order position or not. To describe the width change in the casting process, for each order position j the width at the end of casting wd_j has to be determined. The batchwise supply of material is represented by the binary decision variable z_j, which indicates whether a new charge of material is added before order position j or not. Due to setups, excess material will be lost. For modelling reasons, the setup due to an alloy change is not considered explicitly. The binary decision variable Δ_j indicates a machine setup before order position j which increases the scrap, but not necessarily leads to a stop of production. The binary decision variable γ_j indicates a setup due to a stepwise width

change before order position j which requires a stop of production and increases the scrap as well. The continuous decision variable $scrap_j$ quantifies the cumulated scrap which occurred until the last machine setup before order position j. The continuous decision variable $unused_j$ determines all unused material, including the scrap, which cumulates until the end of order position j.

Using the introduced notation, the following model is derived:

$$MinZ = c^{unused} \cdot unused_n + c^{setup} \cdot \sum_{j=1}^{n} \gamma_j + c^{Op} \cdot \left[\sum_{j=1}^{n} \sum_{i=1}^{n} (wd_i^{max} \cdot x_{ij}) - wd_j \right] \quad (1)$$

subject to:

$$\sum_{i=1}^{n} wd_i^{min} \cdot x_{ij} \leq wd_j \leq \sum_{i=1}^{n} wd_i^{max} \cdot x_{ij} \quad \forall j = 1,...,n \quad (2)$$

$$\sum_{i=1}^{n} wd_i^{min} \cdot x_{ij} - M \cdot \gamma_j \leq wd_{j-1} \leq \sum_{i=1}^{n} wd_i^{max} \cdot x_{ij} + M \cdot \gamma_j \quad \forall j = 2,...,n \quad (3)$$

$$\sum_{i=1}^{n} x_{ij} = 1 \quad \forall j = 1,...,n \quad (4)$$

$$\sum_{j=1}^{n} x_{ij} = 1 \quad \forall i = 1,...,n \quad (5)$$

$$unused_j = unused_{j-1} + CHARGE \cdot z_j - \sum_{i=1}^{n} weight_i \cdot x_{ij} \quad \forall j = 1,...,n \quad (6)$$

$$unused_j \geq scrap_j \quad \forall j = 1,...,n \quad (7)$$

$$scrap_j = scrap_{j-1} \cdot (1 - \Delta_j) + unused_{j-1} \cdot \Delta_j \quad \forall j = 1,...,n \quad (8)$$

$$\Delta_j \geq \sum_{i=1}^{n} \sum_{i'=1}^{n} (1 - fit_{i,i'}) \cdot x_{i,j-1} \cdot x_{i',j} \quad \forall j = 2,...,n \quad (9)$$

$$\Delta_j \geq \gamma_j \quad \forall j = 1,...,n \quad (10)$$

$$z_j \geq \Delta_j \quad \forall j = 1,...,n \quad (11)$$

$$x_{ij}, z_j, \Delta_j, \gamma_j \in \{0;1\} \quad \forall i,j = 1,...,n \quad (12)$$

$$unused_j, wd_j, scrap_j \geq 0 \quad \forall j = 1,...,n \quad (13)$$

$$unused_0 = scrap_0 = 0 \quad (14)$$

In the objective funtion (1) overall cost, consisting of material cost, setup cost and opportunity cost for the loss of production performance, are minimized. Material cost result from the cumulated unused material at the end of the schedule, evaluated with a material cost factor c^{unused}. Setup cost result from machine setups which go along with a stop of production, namely a stepwise width change. They are evaluated with a setup cost factor c^{setup}. Other kinds of setup, namely alloy changes, are evaluated implicit by the resulting scrap. Opportunity cost are approximated through the casting width. The difference between the maximum width of a job on position j and the realized width at position j is evaluated with an opportunity cost factor c^{Op}.

Constraints (2) assure, that the realized casting width at the end of position j fits into the width range of the assigned job. Constraints (3) describe the transition between consecutive jobs regarding the width. If there is no setup, the width at the end of order position $j-1$ has to be within the range of the job scheduled at position j. Each job has to be scheduled at one position (4) and vice versa (5). In (6) the cumulated unused material is updated, similar to holding in other models. If a new charge is added before starting the job at position j, the unused material increases. With each order added to the schedule, the cumlated unused material is reduced by the weight of that order. The amount of cumulated unused material always has to be greater or equal than the cumulated scrap which occurred due to setups until order position j (7). Thus, only material of the current alloy batch can be assigned to orders in the schedule. The cumulated scrap is increased, if some kind of setup occurs directly before position j, identified by Δ_j. Elsewhere, it is left unchanged (8). Constraints (9) and (10) ensure that the setup indicator Δ_j is set to 1 if the alloy or the width is changed due to a setup. Every setup a new material charge needs to be supplied (11). Binary (12) and non-negativity (13) as well as initilization constraints for $unused_0$ and $scrap_0$ (14) complete the formulation.

Due to the quadratic constraints (8) and (9), the presented model formulation is a nonlinear mixed-integer program. Nevertheless, the mentioned constraints are presented as is, to give a better understanding of the models structure. They can be linearized, resulting in a linear mixed integer problem.

4 Case Study

In order to verify our model and to validate its application, a numerical test was conducted. The linearized model formulation was implemented into CPLEX 11.0.

We considered a test set of 15 orders. The order instances are following real world data. For each order, a minimum width between $1,490mm$ and $1,770mm$, a maximum width between $1,540mm$ and $1,950mm$ with a feasible width range between $30mm$ up to $200mm$ was given. The order weights are defined between $18to$ to $30to$. For each order one alloy out of two was assigned. The charge weight was set to $100to$. Finally, the cost factors $c^{scrap} = 500EUR/to$, $c^{setup} = 10,000EUR$ and $c^{Op} = 50EUR/mm$ were assumed.

To check the model for the solution time, 10 test instances with 5, 8, 10 and 12 orders from the test set were solved on a 2.6 GHz CPU with 1GB RAM. For instances with up to 10 jobs, the problem could be solved in reasonable time. The mean computing time increased from $1.5s$ with $\sigma = 0.34s$ for instances with 5 jobs to $1,861s$ with $\sigma = 1,476s$ for instances with 10 jobs. For instances with 12 jobs, only one solution was obtained after $18,900s$. 9 calculation runs were aborted after 24 hours of computation.

In Figure 1 the optimal solution of a test instance with 12 jobs is given. In this example both types of setup can be seen. Between orders 3 and 6, a setup with stop of production is necessary as a result of width constraints. Between orders 2 and 7 a

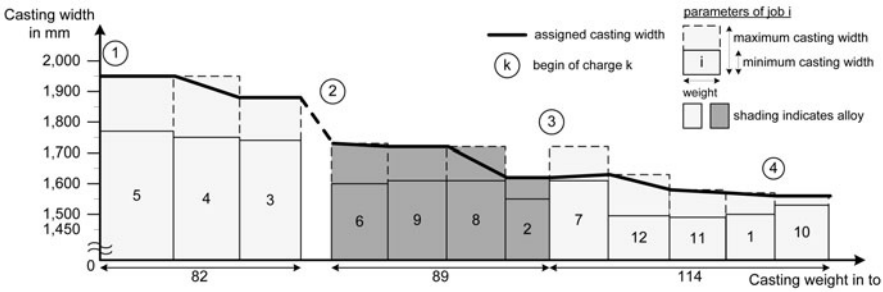


Fig. 1 Optimal solution schedule for problem instance with 12 orders

setup due to an alloy change is necessary. Both setups initiate a new charge of steel. A charge of the same alloy is added without any setup after order 1. Further, the width during the casting process is set as high as possible. However, due to width adjustment during production, it is not always possible to fully utilize each job's maximum width. The reason can be found in the width range of enclosing orders.

5 Conclusion

In this paper we developed a new model for scheduling production orders with flexible job specifications at a continuous casters. We introduced the characteristics of a width range as well as charge related sequence constraints into a new sequencing approach. Thus, assumptions of classic approaches were enhanced to a more realistic aspect of the planning situation. A linearized problem instance was implemented to verify the planning approach. A first numerical study reveals, that for small test instances optimal solutions could be obtained. Further research on adequate solution methods is necessary in order to solve practical problem sizes.

References

1. Milind Dawande, Jayant Kalagnanam, Ho Soo Lee, Chandra Reddy, Stuart Siegel, and Mark Trumbo. The slab-design problem in the steel industry. *Interfaces*, 34(3): 215–225, 2004.
2. A.Drexl and A. Kimms. Lot sizing and scheduling – survey and extensions. *European Journal of Operational Research*, 99: 221–235, 1997.
3. Tieke Li and Yang Xi. An optimal two-stage algorithm for slab design problem with flexible demands. In *Proceedings of the IEEE*, Proceedings of the IEEE, pages 1789–1793, 2007.
4. Stahleisen. *Steel Manual*. Stahleisen-Verlag, Düsseldorf, 2003.
5. Lixin Tang, Jiyin Liu, Aiying Rong, and Zihou Yang. A mathematical programming model for scheduling seelmaking-continuous casting production. *European Journal of Operations Research*, 120: 423–435, 2000.

Optimal Maintenance Scheduling of N-Vehicles with Time-Varying Reward Functions and Constrained Maintenance Decisions

Mohammad M. Aldurgam and Moustafa Elshafei

Abstract In this paper we consider the problem of scheduling the maintenance of fleet of N different vehicles over a given planning horizon T. Each vehicle is assumed to have different time-dependant productivity, mean time to repair, and cost of repair with possible limitations on the maximum number of vehicles that can be repaired at a given time epoch. The objective is to maximize the total productivity minus the cost of repairs. We propose an efficient dynamic programming (DP) approach to the solution of this problem. The constraints are translated into feasible binary assignment patterns. The dynamic programming considers each vehicle as a state which takes one of possible feasible patterns. The algorithm seeks to maximize the objective function subject to specific constrains on the sequence of the selected patterns. An example is given and the DP solution is compared with the best of 50,000 randomly selected feasible assignments. The developed algorithm can also be applied to a factory of N production machines, power generators, or car rental.

1 Introduction and Overview

Consider a company offering a variety of tourism trips. The company has a fleet of N vehicles; each has a specific capacity and time-varying reward function. Each vehicle reward function is assumed to be time-varying depending on the frequency of maintenance activities, where delayed maintenance results in increasing maintenance cost. Customers' demand of trips and vehicle-types are met with the available vehicles. In case of vehicle unavailability, customers are served by rental cars,

Mohammad M. Aldurgam (corresponding author)
King Fahd University of Petroleum and Minerals,Systems Engineering Dept., Dhahran 31261 - KSA, e-mail: aldurgam@kfupm.edu.sa

Moustafa Elshafei
King Fahd University of Petroleum and Minerals,Systems Engineering Dept., Dhahran 31261 - KSA e-mail: elshafei@kfupm.edu.sa

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_60,

which results in a lost profit. With a seasonal demand pattern, the company has put constraints on the number of vehicles that can be maintained during different time epochs. In this paper we formulate a flexible dynamic programming approach, which can accommodate such variety of maintenance constraints, as well as time-dependant productivities and cost of maintenance. Tabari et. al. (2002) used dynamic programming for multi-units electric power generating facilities maintenance scheduling and staffing levels planning using Dynamic Programming. Korpijarvi and Kortelainen (2009) provide a Dynamic program for maintenance planning of electric distribution systems. The model is used to plan maintenance activities options over a given time horizon. Components have a linear reliability function which becomes exponential after a certain threshold value.Time-varying reward can take place due to systems deterioration. Liu et. al. (2001) provide a stochastic dynamic program to schedule preventive maintenance activities for a cutting tool in a flexible manufacturing system.

Aircraft fleet scheduling problem involves crew allocations, assigning aircrafts to flight legs and doing the proper maintenance activities. Usually, integer programming is used, see for example (AhmadBeygi et. al., 2009). El Moudani and Mora-Camino (2000) Considered joint fleet allocation and maintenance scheduling for aircrafts using dynamic programming. Elshafei and Alfares (2008) proposed a dynamic programming algorithm to schedule labor shifts assignments. The model assumes a time-varying cost structure that depends on the work sequence for each employee.

In this paper we propose an efficient dynamic programming algorithm to the solution of fleet maintenance scheduling problem. The constraints are translated into feasible horizontal and vertical binary assignment patterns as in Elshafei and Alfares (2008). The dynamic programming considers each vehicle as a state which takes one of possible feasible patterns. The algorithm seeks to maximize the objective function subject to specific constrains on the sequence of the selected patterns. This paper is organized as follows: the problem definition is provided in section 2. The proposed DP algorithm is presented in section 3. An illustrative example is provided in section 4, and section 5 concludes the paper with the conclusion. Next, a detailed problem definition is provided.

2 Detailed Problem Definition

We consider the problem of maintenance scheduling of N vehicles. The solution will be initially based on the following assumptions:

1. A vehicle, k, can be scheduled for maintenance at most α_k times during the planning horizon.
2. At most r vehicles can be scheduled for maintenance at a given time epoch.
3. Being booked, some vehicles can't be scheduled for maintenance at some time epochs.
4. The normalized productivity of each vehicle is described by a continuous decreas-

ing function with time.
5. Each vehicle, k, has its own maximum possible productivity rate P_k .
6. The productivity of the vehicles is given by $P_k(t) = p_k\beta_k(t)$
7. The maintenance of each vehicle is performed during an entire time epoch, during which the vehicle is not available.
8. The cost of scheduled repair of each vehicle is a continuously increasing function of time between repairs.
9. No major failures of a vehicle during the planning horizon T.

Let, $x_{k,t}$ for $k = 1, 2, \check{E}, N$ and $t = 1, 2, \check{E}, T$ as follows

$$x_{k,t} = \begin{cases} 1 \text{ if vehicle } k \text{ is under service at time } t \\ 0 \text{ otherwise} \end{cases}$$

The schedule of the k^{th} vehicle maintenance may be expressed as: $X_k = \{x_{k,1}, x_{k,2}, \cdots, x_{k,T}\}$ Assumption 1 implies that

$$0 \leq \sum_{t=1}^{T} x_{k,t} \leq \alpha_k, \quad k = 1, 2, \cdots, N \tag{1}$$

While assumption 2 can be written as

$$0 \leq \sum_{k=1}^{N} x_{k,t} \leq r, \quad t = 1, 2, \cdots, T \tag{2}$$

Where, r is the number of vehicles that can be repaired simultaneously in the repair shop. The maintenance assignment of the N vehicles at a time epoch t is given by:

$$v_t = \{x_{1,t}, x_{2,t}, \cdots, x_{N,t}\} \tag{3}$$

The possible N-bit binary representations of (3) will be called vertical patterns.If the workshop can accommodate up to r vehicles at any time, then the number of vertical patterns

$$N_v = \sum_{k=0}^{r} \frac{N!}{k!(N-k)!}$$

Similarly, the schedule of maintenance patterns of a specific vehicle during the horizon T is called a horizontal pattern and is given by

$$h_j = \{x_{j,1}, x_{j,2}, \cdots, x_{j,T}\} \tag{4}$$

Next the proposed Dynamic programming algorithm is provided.

3 The DP Algorithm

We define $J_h(k)$ to be the set of all horizontal patterns for vehicle k, and the size of this set is $N_h(k)$. These $J_h(k)$ patterns are T-bit binary patterns, and we refer to these assignment patterns as the h-patterns (horizontal patterns). The j^{th} h-pattern of the k^{th} vehicle is represented by:

$$h_j^k = \{x_{j,1}^k, x_{j,2}^k, ..., x_{j,T}^k\}$$

The case of a single repair during T implies that the number of patterns is equal to $T+1$, where the no repairs pattern is included. The scheduling solution is represented by an $N \times T$ binary matrix, where the vertical columns belong to the set of vertical patterns $J_v(t)$, and the horizontal rows belong to $J_h(k)$. It should be noticed here that the order of the horizontal rows is irrelevant. In dynamic programming (DP) terminology, the horizontal patterns of a given vehicle are its possible states. The DP will attempt to select one of the states of each vehicle to maximize (or minimize) an overall objective function. Since we need to find N states (one for each vehicle), the DP will proceed in N steps. Since the horizontal order of the vehicles in the assignment table is irrelevant, the DP will consider all the vehicles at any DP step. Let us define:
$h_j^k =$ the current h-pattern number under consideration at the DP step q.
$h_{j1}^{k1} =$ the h-pattern number under consideration at the DP step q -1.
And let $F(q,k,j)$ be the reward (gain) of the DP path ending at DP step q, and horizontal pattern (state) h_j^k. At any given DP step q, the algorithm computes $F(q,k,j)$as follows:

$$F(q,k,j) = max_{k_1,j_1}\{F(q-1,k_1,j_1) + f(q,k,j)\} \quad (5)$$

Where, $f(q,k,j)$ is the local gain of the pattern h_j^k.
$F(q,k,j)$ are calculated recursively. Once we reach $q = N$, the best path is the terminal path which maximizes the objective function among all terminal paths. The selected best path can then be tracked backward to obtain the optimal h-patterns assignment. The selected horizontal patterns can then be converted to their T-bit binary representations, and assembled in the scheduling matrix.

In order to track the optimal path backward, we need to store the path history. Two back tracking matrices $B1(q,k,j)$ and $B2(q,k,j)$ stores the best previous h-pattern indices $k1$, and $j1$ which maximize (5). The algorithm was coded in a MATLAB program. Only the patterns satisfying the constraints are allowed to participate in the optimization of (5). Alternatively, non conforming patterns can be assigned arbitrary large cost or a negative value. The algorithm proceeds according to the following steps:

1. Compute and store the h-patterns and v-patterns
2. Pre-compute the total reward of the horizontal patterns.
3. At DP step q =1, initialize the back tracking matrices $B1(q,k,j)$ and $B2(q,k,j)$, and set the cost $F(1,k,j)$, for $k = 1,2,..N$; and $j = 1,2,\check{E}\, Nh(k)$.
4. For $q = 2,3,\check{E}N$; This is the main dynamic programming loop

5. For $k = 1$ to N, for $j=1$ to $N_h(k)$, for $k1 = 1$ to N and for $j1=1$ to $N_h(k1)$ Check if the sequence $k1$-k, and $j1$-j violate any of the constraints mentioned in the assumption. Set $f(q,k,j)$ to an arbitrary negative value for the violating sequence. Find the optimal $k1$ and $j1$ which optimize (5).
6. Update the back tracking matrix $B1(q,k,j)$ and $B2(q,k,j)$.
7. Continue loop 5
8. At the end of loop 5, $q = N$, find the optimal terminal h-pattern k^*, $j^* = argmax_{j,k} F(N,k,j)$.
9. Use the back tracking matrix $B1(q,k,j)$ and $B2(q,k,j)$, for $q=N-1, N-2, ..2$ to obtain the optimal maintenance h-patterns for all vehicles.

Following is the objective function (J) which maximizes the total reward over horizon T:

$$J = \sum_{k=1}^{N}\sum_{t=1}^{T}(1-x_{k,t})p_k(t)\sum_{k=1}^{N}\sum_{t=1}^{T}x_{k,t}C_k(t) \tag{6}$$

The normalized reliability of the k^{th} vehicle is assumed to be given by

$$\beta_k(t) = \beta_k(t_0)e^{-\frac{(t-t_0)}{\tau_k}} for\, t \geq t_0$$
$$\beta_k(t) = \beta_k^0 e^{-\frac{(t-t_k)}{\tau_k}} for\, t \geq t_k$$

Where $\beta_k(t_0)$ is the initial reliability of the k^{th} vehicle at the start of the horizon T, β_k^0 is the reliability of the vehicle after repair. Hence, the productivity can be defined as follows:

$$P_k(t) = p_k\beta_k(t_0)e^{-\frac{(t-t_0)}{\tau_k}} for\, t \geq t_0$$
$$P_k(t) = p_k\beta_k^0 e^{-\frac{(t-t_k)}{\tau_k}} for\, t \geq t_k$$

The cost of repair of each vehicle can also be estimated as follows:

$$C_k(t) = C_k(t_0)e^{+\frac{(t-t_0)}{\tau_k}} for\, t \geq t_0$$
$$C_k(t) = C_k^0 e^{+\frac{(t-t_k)}{\tau_k}} for\, t \geq t_k$$

After the k^{th} vehicle passes through the scheduled maintenance it returns to a factory condition R_k^0. Next, an illustrative example is provided.

4 Illustrative Example

We consider the case of scheduling the maintenance of 6 cars over 12 time epochs. The following information is available:
P=[10 20 30 40 60 100]; maximum productivity rate of each vehicle in t.

C^0 =[5 5 10 15 20 20]; basic cost of maintenance. R(t0)= [0.99 0.92 0.97 1 0.92 0.8]; initial reliability of the vehicles. R^0 =[1 1 1 1 1 1]; reliability of the vehicle after maintenance. τ = [3 4 2.5 5 3.5 4.5];
We assume only one vehicle can be repaired in a given time period.
The DP optimal assignment as per Eq. (6) is given by:
Vehicle 1: [0 1 0 0 0 0 0 0 0 0 0 0] Vehicle 2: [0 0 0 0 0 0 1 0 0 0 0 0]
Vehicle 3: [0 0 1 0 0 0 0 0 0 0 0 0] Vehicle 4: [0 0 0 0 0 1 0 0 0 0 0 0]
Vehicle 5: [0 0 0 1 0 0 0 0 0 0 0 0] Vehicle 6: [0 0 0 0 1 0 0 0 0 0 0 0]
Total reward = 1,373.8

To validate the result, 50,000 randomly selected feasible solutions were generated, and the objective function (6) was evaluated for each case. The best case was found to achieve total reward =1,352.

5 Conclusion

The paper proposes an efficient DP algorithm for scheduled maintenance of N vehicles, under productivity deterioration and increasing cost of repair with the increased period between successive repairs. The approach can be applied in a variety of constrained and time-varying reward/cost scheduling problems.

Acknowledgements The authors would like to thank King Fahd University of Petroleum and Minerals for its support of this research work.

References

1. S. AhmadBeygi, A. Cohn, and M. Weir. An integer programming approach to generating airline crew pairings. *Computers and Oper. Res.*, 36(4): 1284–1298, 2009.
2. W. El Moudani and F. Mora-Camino. A dynamic approach for aircraft assignment and maintenance scheduling by airlines. *Journal of Air Transport Management*, 6(4): 233–237, 2000.
3. M. Elshafei and H.K. Alfares. A dynamic programming algorithm for days-off scheduling with sequence dependent labor costs. *Journal of Scheduling*, 11(2): 85–93, 2008.
4. J. Korpijavi and J. Kortelainen. A dynamic programming model for maintenance of electric distribution system. *World Academy of Science, Engineering and Technology*, 53: 636–639, 2009.
5. P.H. Liu, V. Makis, and A.K.S. Jardine. Scheduling of the optimal tool replacement times in a flexible manufacturing system. *IIE Transactions (Institute of Industrial Engineers)*, 33(6): 487–495, 2001.
6. N.M. Tabari, A.M. Ranjbar, and N. Sadati. Promoting the optimal maintenance schedule of generating facilities in open systems. In *Proceedings of the 2002 International Conference on Power System Technology, PowerCon 2002, Oct. 2002*, Volume 1, pages 641–645, 2002.

Alternative Quantitative Approaches for Designing Modular Services: A Comparative Analysis of Steward's Partitioning and Tearing Approach

Hans Corsten, Ralf Gössinger, and Hagen Salewski

1 Problem

Modularisation is a possibility to design work sharing services which gains increasing attention in the field of service research (cf. [2, pp. 221]; [3, pp. 811]). The basic concept is to merge single tasks of the underlying service into service modules in such a way, that costs are minimised and that the service supplier is simultaneously enabled to serve a heterogeneous spectrum of customer needs. Because of the integrated nature of the co-production of services by service supplier and customers, the relevant costs are coordination costs[1], which arise when interdependent tasks are to be fulfilled by different actors, or when these tasks are not completely complementary. In the context of service production, tasks can be combined to service modules up to such a module size, that existing interdependencies are still minimal in terms of coordination costs. Higher coordination costs only occur, if interdependent tasks are assigned to different modules and when the activities, which different actors are responsible for, need do be adjusted.

To support the decisions in the design of service modules, different quantitative approaches are suggested (cf. [5, pp. 1675]; [6, pp. 40]), which are based on the model of Design Structure Matrices (DSM). On the basis of DSM, partitioning algorithms can be utilized to identify modules. A technique, which is often used, is the partitioning and tearing approach described by D. V. Steward. However, the performance of this approach has not been examined in adequate detail. This contribution aims at a test-based evaluation of Steward's approach. The performance of Stew-

Hans Corsten and Hagen Salewski
Dept of Business Administration and Production Management, University of Kaiserslautern, Gottlieb-Daimler-Straße 42, 67663 Kaiserslautern, Germany, e-mail: corsten@wiwi.uni-kl.de, salewski@wiwi.uni-kl.de

Ralf Gössinger
Dept of Business Administration, Production and Logistics, University of Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany, e-mail: Ralf.Goessinger@tu-dortmund.de

[1] The case, that single tasks cause direct costs of work sharing, is not considered.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_61,

ard's approach is compared with the partitional clustering approach by J. Reichardt, which aims at the detection of structures in large networks. Both approaches need to be modified to solve the presented problem. To evaluate the performance, the standard criteria solution time and solution quality are used. Additionally, the influences of the DSM's composition are considered and the effects of size, structure, and density of the matrix are examined.

2 Approaches

Modularisation is realised by the assignment $x_{v.m}$ $(x_{v.m} \in \{0,1\})$ of tasks v $(v = 1,...,V)$ for a specified service path[2] to modules m $(m = 1,...,M; M < V)$. The assignment should ensure that

- the relevant coordination costs between ordered modules, which are provided for each task combination v and v' as elements $a_{v.v'} \geq 0$ of the DSM $\mathbf{A}[V \times V]$, are minimised; the order of the modules is enforced by the range of the indices m and m' in the objective function,
- the number of tasks in one module cannot exceed the coordination capacity of an actor and each module must exceed a given number of tasks which represent the minimum capacity utilisation of this actor, and
- the set of all tasks must be decomposed completely into disjoint modules.

These requirements can be modelled as a quadratic assignment problem with $V \cdot M$ decision variables and $2 \cdot M + V$ constraints.

$$\text{Min! } K = \sum_{m=1}^{M-1} \sum_{m'=m+1}^{M} \sum_{v=1}^{V} \sum_{v'=1}^{V} x_{v.m} \cdot x_{v'.m'} \cdot a_{v.v'}$$

$$\text{s.t.: } \sum_{v=1}^{V} x_{v.m} \leq S^{max} \qquad \forall m$$

$$\sum_{v=1}^{V} x_{v.m} \geq S^{min} \qquad \forall m$$

$$\sum_{m=1}^{M} x_{v.m} = 1 \qquad \forall v$$

$$x_{v.m} \in \{0,1\} \qquad \forall v, m$$

The optimal solution of the model can be found by applying exact methods for small DSM sizes only.

[2] Since only specified service paths are considered, the combination of tasks of a path can only be partial services. A construction of service modules has to be made across all paths. However, this distinction is not kept up in this paper and the term *module* is used instead.

Steward's approach (cf. [6, pp. 40]) is a two-step heuristic designed for binary DSMs. In the first step, partitioning, interdependent service tasks are combined in such a way, that

- only one-directional dependencies remain between modules,
- modules contain as few tasks as possible, and
- modules are in an order which minimises the total number of dependencies to preceding service modules.

After partitioning, the internal structure of the resulting modules is analysed, and tearing of internal dependencies is used to create an internal order of the service tasks. The tearing described by Steward leads to unfavourable small modules. For this reason, the tearing was enhanced a by a system grid analysis (c.f. [1, pp. 176]).

Reichardt's approach (cf. [4, pp. 31]) assigns the service tasks to an ex ante unknown number of service modules, based on the existence of dependencies between tasks. The approach uses a strongly modified simulated annealing algorithm. The underlying analogy is the cooling of ferromagnetic spin glass, which leads to a specific energy function and a specific acceptance probability. This approach was also modified to ensure the required maximum and minimum module sizes.

3 Results

The performance of the approaches is tested for different matrix sizes with fixed module sizes: 8×8 (2 tasks), 16×16 (4), 32×32 (4), 64×64 (8), and 128×128 (16). The densities[3] of the modules and of the overall DSM vary from 20% to 90% for the modules and 10% to 50% for the DSM.

To improve the visualisation of results, both density values are combined to a contrast ratio:

$$\kappa = \rho_{Module} \cdot \frac{n_{DSM} - n_{Module}}{\tilde{n}^{+}_{DSM} - \tilde{n}^{+}_{Module}} \quad \text{, with:}$$

ρ = density of the DSM or the to-be-found modules,
n = number of all elements of the DSM and, accordingly, the number of elements in all to-be-found modules, and
$\tilde{n}^{+}$ = rounded product of the number of elements n and the corresponding density value.

For each combination of parameters, 10 different DSMs are generated randomly. Of all possible combinations, only those represent the given problem, for which the module density is higher than the matrix density. Yet, combinations are used within the described ranges where the DSM density is up to one step above the module density. The total number of tested problems is 1,860.

[3] The density is the relation of the number of elements that equal one and the number of all elements inside of a DSM or a specified part of it.

For each problem, the value of the objective function, the runtime, and the modularity Q are observed, where Q is defined as

$$Q = \left(1 - \frac{\text{number of dependencies in the upper triangular matrix}}{\text{number of all dependencies}}\right)$$

in an ordered DSM according to the resulting modularisation. The maximum modularity ($Q = 1$) occurs if no interdependencies exist in the upper triangular matrix.

The test system used was a PC running on Intel core i7 processors with 2.8 GHz, 8 GB of RAM and Windows 7 64-bit. To compute the optimal solution we used FICO Xpress Optimization Suite 7.0.2. The heuristics were implemented using Matlab R2010a, not using parallel computing techniques.

The left half of figure 1 shows the typical course of the objective value, the modularity, and the runtime of the two heuristic approaches and the exact approach for an 8×8 -DSM. In these figures, the observed values are plotted against the contrast for a module density of 60%. Solving the decision model described in section 2 exactly causes a very high computational effort and, thus, the worst results concerning runtime. It can be seen that Reichardt's approach achieves a better solution quality than Steward's, whereas Steward's approach is faster. This is due to the fact that the combination of the DSM density of 10% and the module density of 90% is one of the above mentioned extreme cases, and the last point plotted is for a DSM density of 20%. This can also be seen in the plot at the bottom of figure 1, where the objective value for the highest contrast is greater than zero.

For larger-sized matrices, the runtime of the exact approach is too high to generate good solutions in acceptable time. Because the progression against matrix size for the three observed values is similar for different density combinations, the DSM density of 30% and the module density of 60% are chosen for a typical representation[4]. A first and expected observation is the general decrease of modularity and the increase of objective values. The total number of interdependencies increases with an increasing size of the DSM at set dependencies.

It is a remarkable observation, that for the parameter combination plotted in the right half of figure 1 the objective values of Steward's approach are better than the values of Reichardt's approach for matrix sizes of 32 and 64 service tasks. The difference of the objective values between both approaches decreases with increasing matrix sizes. However, this observation cannot be generalised for all combinations of matrix sizes and densities. The conjecture, that the quality of the results of Steward's and Reichardt's approaches level with growing matrix size, deserves additional studying. Reichardt's approach starts off with a better modularity then Steward's, but Steward's approach generates better results concerning the modularity for matrix sizes of 32 single service tasks and above. The runtime of both approaches increases with the matrix size, yet the runtime of Steward's approach rises stronger and exceeds the runtime of Reichardt's approach starting at a matrix size of 64 service tasks.

[4] The chosen combination of densities corresponds to the contrast of $\kappa \approx 2.2$, which is also marked at the plot for the results of the 8×8 -matrices on the left.

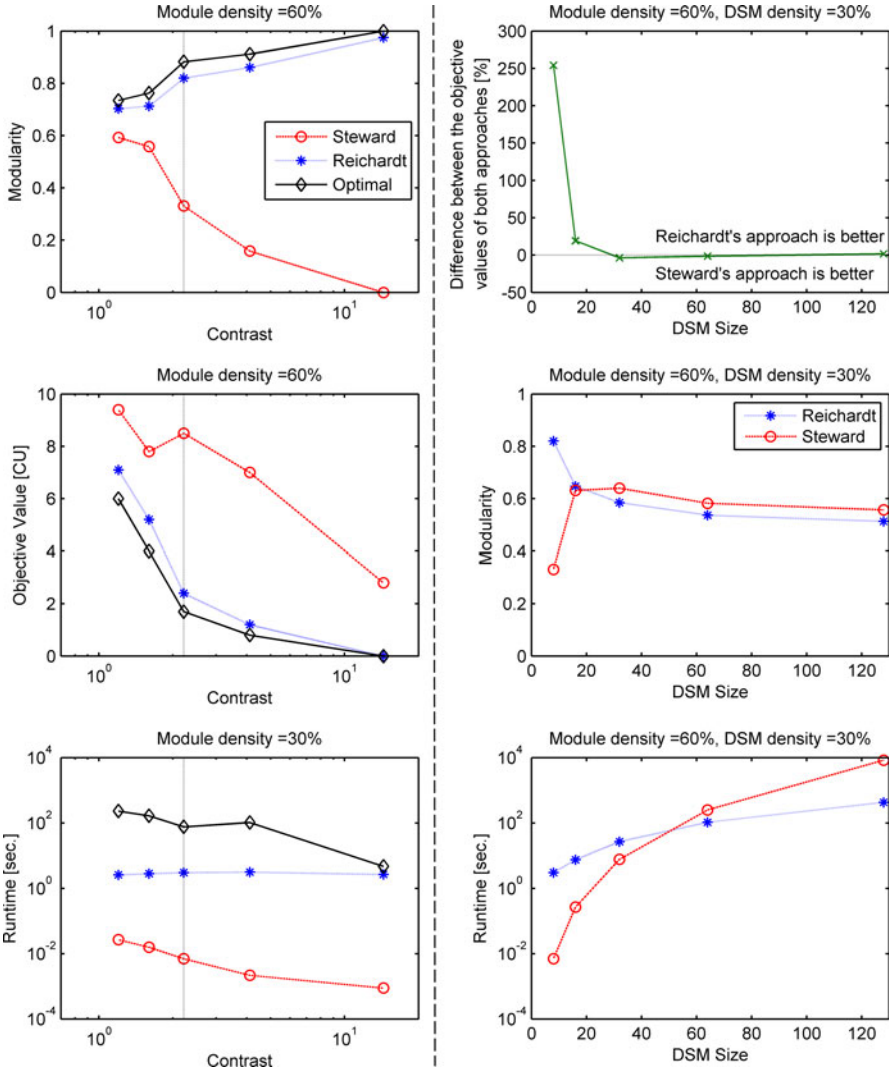


Fig. 1 Modularity, objective value and runtime of Steward's and Reichardt's approaches as well as for the optimal solution for different densities in an 8×8 -DSM (left half) and of Steward's and Reichardt's approach for different DSM sizes (right half)

4 Concluding Remarks

The presented paper evaluates the capability of Steward's approach for the specific problem of finding an assignment of service tasks to service modules for a given service path which minimises the coordination costs. It could be shown that, for small problem sizes, the solution quality of Steward's approach is dominated by Reichardt's approach concerning all tested parameter combinations. For larger matrices, starting at the dimensions of 32×32, the qualities of the results converge. Further research, involving statistic testing, needs to be conducted to judge the quality more accurately. Since the presented problem belongs to the area of tactical planning, the observed runtime is acceptable for both approaches. For further assessment of Steward's approach, our future research will focus on a set of questions beyond the scope of this paper: How is the quality of the results of Steward's approach affected, if

- the used test matrices are generated with different sizes of to-be-found modules inside one matrix,
- the size of to-be-found modules exceeds the given maximum size of the modules and vice versa,
- the densities do not only vary concerning the to-be-found modules, but also concerning other selected parts of the DSM,
- instead of binary DSMs, the general case of interdependencies with different strengths is considered[5].

References

1. H. Corsten and R. Gössinger. Modularisierung von Dienstleistungen: Untersucht am Beispiel von Logistikdienstleistungen. In H. J. Gouthier, C. Coenen, H. S. Schulze, and C. Wegmann, editors, *Service Excellence als Impulsgeber*, pages 163–185. Wiesbaden, 2007.
2. R. Di Mascio. Service Process Control: Conceptualising a Service as a Feedback Control System. *Journal of Process Control*, 12(2): 221–232, 2002.
3. J. Jiao, Q. Ma, and M. M. Tseng. Towards High Value-Added Products and Services: Mass Customization and Beyond. *Technovation*, 23(10): 809–821, 2003.
4. J. Reichardt. *Structure in Complex Networks*, volume 766 of *Lecture Notes in Physics*. Berlin, Heidelberg, 2009.
5. M. E. Sosa, S. D. Eppinger, and C. M. Rowles. The Misalignment of Product Architecture and Organizational Structure in Complex Product Development. *Management Science*, 50: 1674–1689, 2004.
6. D. V. Steward. Systems Analysis and Management: Structure, Strategy and Design. New York, 1981b.

[5] To address this question, the Steward approach needs to be further modified, since Steward only describes applications to binary matrices.

Applying DEA Models for Benchmarking Service Productivity in Logistics

Marion Steven and Anja Egbers

Abstract This paper contrasts two extensions of Data Envelopment Analysis (DEA) under constant returns to scale (CRS) assumption for benchmarking service productivity in logistics. The applicability of an additive and a multiplicative Integrated CCR model is evaluated by means of an application.

1 Introduction

In manufacturing-oriented concepts, productivity is defined as ratio of outputs achieved and inputs used. As high productivity is required in every division of an enterprise, this input and output orientation also concerns logistic activities [8]. Because of intangible outputs, concurrent production and consumption processes and customer participation in those processes, it is difficult to transfer the concept of productivity without amendments. Given that productivity measurements in services are normally partial measurements, it is impossible to define "one unit of a [logistic] service" [6].

After an overview of the constitutive attributes of logistic services (section 2), the paper deals with the question how productivity of logistic services is measured (section 3). Furthermore, productivity of logistic services will be evaluated by Data Envelopment Analysis (DEA) considering constant returns to scale (CRS) which was developed by Charnes, Cooper and Rhodes [2]. In order to analyse the productivity of services, DEA is extended by additive and on the other hand by multiplicative composition (section 4). The applicability of the approaches proposed is illustrated with an application in section 4.

Marion Steven, Anja Egbers
Ruhr-Universität Bochum, Lehrstuhl für Produktionswirtschaft, Universitätsstrasse 150, Bochum
e-mail: marion.steven@ruhr-uni-bochum.de, anja.egbers@ruhr-uni-bochum.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_62,

2 Constitutive Attributes of Logistic Services

During the service production process, inputs are transformed into economic outputs through factor combination. In the first step the internal inputs are combined to provide the infrastructure for service production. By using this infrastructure in the production process, the characteristics of a given external factor are transformed in the second step [4]. According to the definitions of service production, logistic services may be regarded from input-, process- and output-oriented perspectives [10, 8]:

- *Input-oriented definitions* concentrate on the promise of certain logistic services and the allocation of equipment.
- In contrast, *process-oriented definitions* focus on the performance of service delivery which is achieved by the combination of internal and external factors.
- Finally, *output-oriented definitions* refer to the effects of logistic services generated for the customer.

The basic logistic processes such as transportation, handling and stocking are the typical *primary logistic services* and present the object for the following analysis. In contrast, so-called *secondary logistic services* such as marketing activities or after-sale services will be excluded from analysis, as their outputs are intangible [6] and thus output values incommensurable.

3 Measurement of Logistic Service Productivity

Although it is often useful to measure partial productivity, only total productivity measurements can give complete information about the efficiency of services [6]. Whereas real assets like materials may be stockpiled until consumption, it is impossible to stock logistic services in periods of low demand and use them in periods of high demand. Regarding this, as shown in figure 1, productivity of logistic services such as transport processes may be defined by an ex ante productivity ratio (internal production of services) and is different from an ex post definition, including the influence of an external factor, while considering the effect of demand [4, 6]. Concerning these service characteristics we can distinguish between an ex ante and an ex post productivity. The ratio of outputs (y^a), like transport capacity, and inputs (x^a), like both staff or number of vehicles, builds the ex ante productivity. The ex post productivity is defined as ratio of outputs (y^b), such as transported goods, and inputs (x^b), like working hours of staff or fuel consumption. The ex ante productivity focuses on internal processes of capacity provision while the ex post productivity concentrates on the transformation of the capacity and other input factors into marketable services.

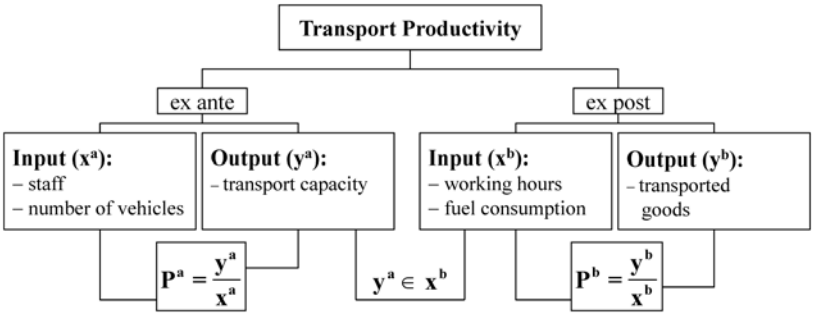


Fig. 1 Transport Productivity

4 Two Data Envelopment Analysis (DEA) Approaches

With DEA it is possible to measure the relative efficiency of "Decision Making Units"'(DMUs). A DMU may be regarded as a collection of enterprises, divisions or administrative units with common inputs and outputs [2]. For analysing the productivity of transport processes it is advantegeous to measure the ex ante and ex post productivity simultaneously. Therefore, the following subsections deal with two extensions, considering the Integrated CCR model developed by Chiou et al.[3]. In the additive Integrated CCR model (cf. 4.1) ex ante and ex post productivity are balanced equally, while in multiplicative models (cf. 4.2) the lower value has an wider influence on the total productivity.

4.1 Additive Integrated CCR Model

The additive Integrated CCR model in the CRS context considers both ex ante and ex post productivity and is formulated as follows [3]:

$$\max_{v^a,u^a,v^b,u^b} P_o^{add} = \left(\frac{\sum_{j=1}^{J} u_j^a y_{jo}^a}{\sum_{i=1}^{I} v_i^a x_{io}^a}\right) + \left(\frac{\sum_{s=1}^{S} u_s^b y_{so}^b}{\sum_{r=1}^{R} v_r^b x_{ro}^b}\right) \tag{1}$$

$$s.t. \quad \left(\frac{\sum_{j=1}^{J} u_j^a y_{jk}^a}{\sum_{i=1}^{I} v_i^a x_{ik}^a}\right) \leq 1, \qquad k \in K,$$

$$\left(\frac{\sum_{s=1}^{S} u_s^b y_{sk}^b}{\sum_{r=1}^{R} v_r^b y_{rk}^b}\right) \leq 1, \qquad k \in K,$$

$$v_i^a \geq 0, u_j^a \geq 0, v_r^b \geq 0, u_s^b \geq 0, \qquad i \in I, j \in J, r \in R, s \in S.$$

This model aims to maximize the ex ante and ex post productivity $P_o \in [0,2]$ of a transport process o by simultaneously computing the virtual multipliers v_i^a (v_r^b)

corresponding to the i th ex ante input x^a_{io} (r th ex post input x^b_{ro}) and the virtual multipliers u^a_j (u^b_s) corresponding to the j th ex ante output y^a_{jo} (s th ex post output y^b_{so}). The first and second constraints ensure that there is no other transport process k with an efficiency of more than 100% considering the same multiplier values for each transport process. Supplemental constraints indicate positive multipliers. Based on the Charnes-Cooper transformation [1], it is possible to reformulate this model as a standard linear program [9]. Furthermore, DEA models can be distinguished whether they are input-oriented or output-oriented. Input-oriented models aim to minimize the input values considering fixed outputs. On the other hand, output-oriented models fix inputs while the outputs are maximized [5].

4.2 Multiplicative Integrated CCR Model

In contrast to the first model, the objective function in this proposed model evaluates the maximum ex ante and ex post transport productivity $P_o \in [0,1]$ of process o by multiplicative composition. While the objective function differs from model (1), the constraints are the same as in additive Integrated CCR-model [3]:

$$\max_{v^a,u^a,v^b,u^b} P_o^{mult} = \left(\frac{\sum_{j=1}^{J} u^a_j y^a_{jo}}{\sum_{i=1}^{I} v^a_i x^a_{io}}\right) * \left(\frac{\sum_{s=1}^{S} u^b_s y^b_{so}}{\sum_{r=1}^{R} v^b_r x^b_{ro}}\right) \tag{2}$$

Therefore, both the ex ante and ex post productivity of every analysed transport process k do not exceed one and all multipliers are positive. As in model (1), x^a_{io} and y^a_{jo} describe the i th ex ante input and j th ex ante output of the transport process o , while x^b_{ro} and y^b_{so} express the r th ex post input and s th ex post output respectively. By simultaneously solving the corresponding virtual input multipliers v^a_i (v^b_r) and the virtual output multipliers u^a_j (u^b_s), the productivity of transport process o is maximal and equals one.

5 Application

In this analysis the productivity of transport processes of the eight biggest airlines worldwide (relating to the number of passengers) will be benchmarked. Table 1 contains adapted input and output data available from the annual World Air Transport Statistics published by the International Air Transport Association (2008). First of all, both fleet size of the airlines and staff, like pilots and flight attendants, constitute two input variables x^a_{ik} of ex ante productivity, while the output variable y^a_{jk} is the available tonne kilometer. By this ratio, the productivity of allocation, resp. the availability of a transport process, will be measured. Beyond that, for analysing the ex post productivity, two output variables y^b_{sk} (number of passengers and freight

tonnes) as well as two input variables x^b_{rk} (available tonne kilometer and flight hours) are considered. The transport capacity provided is specified as available tonne kilometer and constitutes the output variable of ex ante as well as the input variable of ex post productivity.

Table 1 Summary of input and output data for the case of eight airlines *

Airline	fleet size	staff	tkm available	flight hours	passengers	freight tonnes
	x^a_{1k}	x^a_{2k}	y^a_{1k}, x^b_{1k}	x^b_{2k}	y^b_{1k}	y^b_{2k}
Southwest Airlines	636	729.885	22.042.728	2.041.621	11.840.516	148.486
American Airlines	655	35.655	38.978.027	2.079.938	20.214.124	508.602
Delta Airlines	446	24.204	28.823.939	1.758.128	14.983.614	313.508
United Airlines	460	27.096	37.786.202	1.612.446	19.504.140	1.809.439
US Airways	360	10.149	11.312.310	942.363	6.361.855	72.713
Lufthansa	417	41.076	28.122.022	1.474.305	12.305.975	1.237.249
Air France	387	17.545	39.949.365	1.915.077	19.051.855	1.445.183
Northwest Airlines	398	15.051	21.024.590	1.273.766	10.663.956	656.246

* [7]: World Air Transport Statistics.

Analysing the results of DEA denoted in table 2, US Airways is benchmarked as most productive by the proposed Integrative CCR model with additive as well as with multiplicative composition [3]. The additive productivity ($P^{add}_o = 2$) is equal to the sum of scores of ex ante (P^a_o) and ex post (P^b_o) productivity. The maximum multiplicative productivity value P^{mult}_o equals one. As can be seen additionally, the ex post productivity P^b_o of United Airlines is equal to 1 , while the ex ante productivity P^a_o is 0,7887. For this reason, United Airlines is required to reduce their fleet size or staff to 78,87 % in order to be efficient to produce the available tonne kilometer. The second and third best airlines in additive Integrated CCR model are United Airlines (1,7887) and Delta Airlines (1,7831), while in the multiplicative model the sequence is reciprocal (Delta Airlines= 0,7890 and United Airlines= 0,7887).

This shows that the total productivity value is highly dependent on the applied model. The order of DMUs depends on the model chosen, which is subject to the preferences of the decision maker. The abovementioned additive Integrated CCR model implicates equal weights for ex ante and ex post productivity, while in the multiplicative Integrated CCR model the total productivity is dependent on the lower productivity value. The assumption of equal weights in additive Integrated CCR model can be generalized by an additional model considering unequal weights [3].

Table 2 Results from Integrative CCR model *

Airline	P_o^a	P_o^b	P_o^{add}	P_o^{mult}
Southwest Airlines	0,6229	0,9563	1,5792	0,5957
American Airlines	0,5955	0,9816	1,5771	0,5845
Delta Airlines	0,8150	0,9680	1,7831	0,7890
United Airlines	0,7887	1,0000	1,7887	0,7887
US Airways	1,0000	1,0000	2,0000	1,0000
Lufthansa	0,8300	0,9188	1,7487	0,7625
Air France	0,7615	0,9133	1,6748	0,6955
Northwest Airlines	0,5996	0,9516	1,5512	0,5706

* Solved by DEAFrontier™ Free Trial Version

6 Conclusion

This paper presented two different approaches to benchmark productivity of logistic services. The additive Integrated CCR model aims to maximize the ex ante and ex post productivity of logistic services equally, while the productivity in the multiplicative Integrated CCR model is highly dependent on the lower value of ex ante respectively ex post productivity.

References

1. A. Charnes and W.W. Cooper. *Management Models and Industrial Applications of Linear Programming*. John Wiley and Sons, New York, 1961.
2. A. Charnes,W.W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2: 429–444, 1978.
3. Y.-C. Chiou, L.W. Lan, and B.T.H. Yen. A joint measurement of efficiency and effectiveness for non-storable commodities: Integrated data envelopment analysis approaches. *European Journal of Operations Research*, 201: 477–489, 2010.
4. H. Corsten and R. Gössinger. *Dienstleistungsmanagement*. Oldenbourg, München, Wien, 2007.
5. K. Cullinane, T.-F. Wang, D.-W. Song, and P. Ji. The technical efficiency of container ports: Comparing data envelopment analysis and stochastic frontier analysis. *Transportation Research Part A*, 40: 354–374, 2006.
6. C. Grönroos and K. Ojasalo. Service productivity – Towards a conceptualization of the transformation of inputs into economic results in services. *Journal of Business Research*, 57: 414–423, 2004.
7. International Air Transport Association. World Air Transport Statistics. 2008.
8. H.-C. Pfohl. *Logistiksysteme – Betriebswirtschaftliche Grundlagen*. Springer, Berlin, 2010.
9. H.D. Shermann and J. Zhu. *Service Productivity Management — Improving Service Performance using Data Envelopment Analysis (DEA)*. Springer, New York, 2007.
10. M. Steven. *Handbuch Produktion, Theorie – Management – Logistik – Controlling*. Kohlhammer, Stuttgart, 2007.

Quality Performance Modeling in a Deteriorating Production System with Partially Available Inspection

Israel Tirkel and Gad Rabinowitz

Abstract This work studies quality output of a production system applying simulation and analytical models. It originates in semiconductors, but can be adapted to other industries. We investigate the impact of Inspection Policies (IP) on Flow-Time (FT) and quality, as measured by Out-Of-Control (OOC). Results indicate that growing inspection rate reduces OOC and increases FT until OOC is minimized, then OOC starts to grow while FT continues to increase. Dynamic IP's are superior to the known static IP. Maximum inspection rate or inspection utilization does not acheive minimum OOC. Operation and repair times variability affect OOC more than their statistical functions.

1 Introduction

Operations struggle with the tradeoff between quality and Flow-Time (FT) due to their significant impact on profit. High quality increases revenue and short FT reduces cost. Our study originates in semiconductors manufacturing, but can be adapted to other process-manufacturing industries. Mittal and McNally [9] discuss the motivation in wafer fabrication which heavily relies on in-line inspection. We investigate the impact of Inspection Policies (IP), tool availability and utilization on FT and quality, in a random deteriorating production system.

Studies of semiconductors operations frequently apply queueing models to production lines [4]. FT versus quality investigation is not common and rely on the known FT-Yield premise [8], or challenge its authenticity [7]. Deteriorating production systems have been studied since the early work of Duncan [2]. Later work

Israel Tirkel
Ben-Gurion University of the Negev, Beer-Sheva, Israel e-mail: tirkel@bgu.ac.il

Gad Rabinowitz
Ben-Gurion University of the Negev, Beer-Sheva, Israel e-mail: rgadi@bgu.ac.il

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_63,

focused on cost and deterioration functions [5]. Maintenance of systems has also been studied extensively [14]. Additional studies that investigate operation and repair time functions conclude that Exponential is appropriate for operation [1, 3] and repair [3, 12], and Lognormal appropriate for repair [12]. Semiconductors inspection studies [6] applied Exponential function for operation and repair times.

Recent work [13] established a Yield-FT model in wafer fabrication. This paper extends previous studies by: (*i*) generalizing Yield to Out-Of-Control (OOC) ratio, (*ii*) considering inspection with partial availability, and (*iii*) associating inspection utilization with quality output. It is done under evaluation of static and dynamic IP's using simulation and analytical models. Section 2 describes the deteriorating production system model, Section 3 defines the variables and scenarios, Section 4 presents the results obtained, and Section 5 infers conclusions.

2 The Model

This section describes the production system model. Fig. 1 represents a repetitive building block in a production line, consisting of two consecutive production steps and a single inspection step. The Feeder feeds processed items into the Bleeder. Items departing the Feeder can either continue directly to the Bleeder (Path 1), or initially be routed to the Metro and then to the Bleeder (Path 2). The Metro inspects some of the items in order to determine the Feeder's state; we assume items are not reworked or discarded, and do not acquire flaws in Metro.

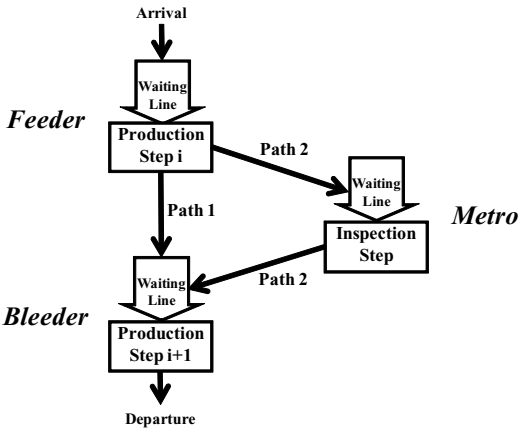


Fig. 1 Three-step production system

We further assume that each component behaves as a single queue with a single waiting-line and a single server (FIFO). The Feeder is M/M/1, the Metro is E/M/1 or G/M/1 (when no repair), and the Bleeder is G/M/1. Inspection involves longer

FT of items processed in the Metro, some impact on the Bleeder's FT and none on the Feeder's. Production scenarios are defined by: (*i*) system's arrival rate (λ), (*ii*) Feeder, Metro and Bleeder service rates (μ_F , μ_M , μ_B), and (*iii*) Metro availability function. We assume that the steps in the production line are independent in their quality performance. The purpose of in-line inspection is to assess if an item's quality is OOC or In-Control (IC), and determine the processing tool state accordingly. In case of OOC, item is not discarded since it is not necessarily unsalable (e.g. out of specification limits). The Feeder's states are defined as either IC or OOC. The state is determined at the instance process starts: stay IC at a known transition probability p , transfer from IC to OOC at probability $1-p$, stay OOC at probability 1 . Once an inspection is complete, the state is recognized with no errors; if OOC is detected, the Feeder is corrected to IC at the first instance it is not processing.

3 Variables and Scenarios

This section defines key variables, performance measures, IP's and scenarios applied. The Measure Cycle (MC) is the number of consecutive items processed between two inspected ones; the IP determines the distribution of MC. Inspection Time (IT) is measured since an item departs the Feeder and until it departs the Metro. During the IT the Feeder continues to process items; longer IT results in a larger number of processed items and consequently with higher average OOC rate. The expected FT of m items processed in a step is defined by,

$$FT = \frac{1}{m}\sum_{j=1}^{m}(CompleteTime_j - StartTime_j) \tag{1}$$

The OOC ratio of m items processed in a step is defined by,

$$OOC = \frac{Quantity\ of\ OOC\ items}{m\ Total\ Incoming\ Items} \tag{2}$$

The tool(s) availability in a step is defined by,

$$Av = \frac{MTBF}{MTBF + MTTR} \tag{3}$$

where, $MTBF$ is Mean Time Between Failures and $MTTR$ is Mean Time To Repair. The tool(s) utilization in a step is defined by,

$$Util = \frac{TotalProcessingTime}{TotalTime} \tag{4}$$

where, $TotalProcessingTime = MTBF + MTTR$.

IP's are decision algorithms associated with parameters, classified [11] as (*i*) Predictive (Static), where in-advance schedule of items for inspection is generated (# 1

below), or (*ii*) Completely-Reactive (Dynamic), where no schedule is generated and decisions are made in response to real-time events (# 2, # 3 below):

1. FMR (Fixed Measure Rate); if $x-1$ consecutive items depart the Feeder, send the x^{th} item to inspection. This IP is common practice in wafer fabs.
2. MLT & BMT (Metro Less Than & Bleeder Less Than); if Metro is Less Than x items AND Bleeder is More Than y items, send item to inspection.
3. Freshest (Freshest Item); if Metro is empty, send the Freshest Item to inspection, where Freshest Item is the most recent item processed by the Feeder.

Moore [10] projected that "the number of transistors on a chip will double every two years", known as Moore's Law. One of its effects is lengthier production lines, currently reaching 600 steps with 45% inspections. Consequently, FT has increased and quality output improved. Our scenarios are in accordance, and the model reflects a proportion of one inspection per a production step. The input rate $\lambda = 1$ indicates one item per hour arrival rate, where an item is a 25 wafer-lot (18,000 wafers per month is medium size fab). Traffic intensity of production steps is 0.9 (competitive goal). Transition probability $p = 0.99$ triggers OOC every 100^{th} item, on average. MC range is 1-10. Metro operation and repair time are distributed Exponential and Lognormal, at availability levels 80% -100% .

4 Results

This section presents the results obtained using: (*i*) analytical model with FMR IP at 100% Availability, and (*ii*) custom simulation (C-sharp) with all IP's at various Availability levels.

Fig. 2 illustrates OOC ratio vs. FT as a quasi-convex shape, which is typical regardless of scenario or parameters applied. It indicates that growing inspection rate reduces OOC ratio and increases FT until minimum OOC ratio is reached and then starts to grow, while FT continues to increase. The cause for the OOC ratio increase is higher IT due to longer waiting time. Metro Availability levels lower than 100% were generated using Exponential operation and repair times. Clearly, minimum OOC ratio increases with decreasing Availability. Notice that at each Availability level the dynamic IP's are superior to the static IP in achieving minimum OOC ratio; this preference grows as Availability decreases. Based on a similar comparison, Freshest is superior to MLT&BMT IP.

Fig. 3 illustrates OOC ratio vs. Metro Utilization with similarly shaped curves. Minimum OOC ratio is lower for dynamic IP's, explained by achieving higher Utilization at similar average MC. Also, minimum OOC ratio is not achieved when maximizing Metro utilization, as usually targeted by operations management. Clearly lower Availability drives lower Utilization, which decreases faster for static IP than for dynamic IP's, and explains the growing preference of dynamic IP's at lower Availability.

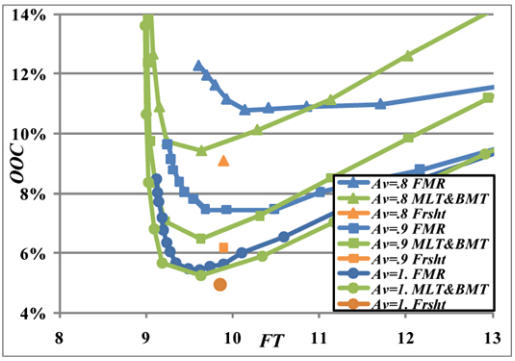


Fig. 2 OOC vs. FT - Comparison of IP's and Availability levels

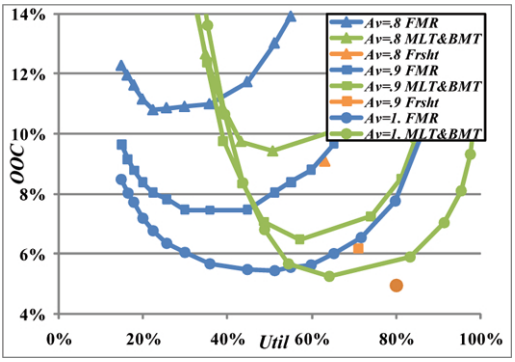


Fig. 3 OOC vs. Util - Comparison of IP's and Availability levels

Identical scenarios were investigated for Metro Availability with Lognormal operation and repair times. OOC results obtained at 90% Availability were slightly lower across all FT-axis (0.25% lower at minimum OOC ratio) and even lower at 80% Availability (0.5% lower at minimum OOC). This indicates that operation and repair functions slightly impact OOC ratio, where Lognormal are favorable over Exponential times. Additional experiments investigated functions of times using lower Coefficient Of Variation (CoV). Results indicated that (*i*) reduced CoV using Lognormal times drive decrease in OOC ratio (e.g. minimum OOC ratio at 80% Availability with FMR achieved 9% with 0.4 CoV vs. 11% with 1.0 CoV), and (*ii*) the gap between OOC ratio reduces when using different functions at the same yet lower CoV (e.g. 80% Availability and 0.4 CoV with FMR was almost identical when Lognormal and Uniform functions were compared). Clearly, this indicates that the impact of operation and repair times CoV is larger than the impact of their functions.

5 Conclusions and Future Research

Dynamic IP's are superior to the common known static IP. Minimum OOC ratio is not achieved at maximum inspection rate or at maximum inspection tool(s) utilization. Dynamic IP's achieve higher tool utilization at minimum OOC ratio. Clearly, higher availability improves quality performance. Operation and repair times variability have a greater effect on OOC ratio than their distribution functions. Lower inspection tool availability also drives reduced tool utilization. The superiority of dynamic over static IP's is greater at lower tool availability and at higher variability of operation and repair times. This work paves the way to study other models of maintenance and raises the challenge to study the integrative impact of other operations management policies on quality and FT.

References

1. P. Chandrasekhar and R. Natarajan. Confidence limits for steady state availability of systems with lognormal operating time and inverse gaussian repair time. *Microelectronics Reliability*, 37(6): 969–971, 1997.
2. A. J. Duncan. The economic design of x-charts used to maintain current control of a process. *Journal of the American Statistical Association*, 51(274): 228–242, 1956.
3. L. Haque and M. J. Armstrong. A survey of the machine interference problem. *European Journal of Operational Research*, 179(2): 469–482, 2007.
4. W. J. Hopp, M. L. Spearman, S. Chayet, K. L. Donohue, and E. S. Gel. Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34: 119–130, 2002.
5. F. Hu and Q. Zong. Optimal production run time for a deteriorating production system under an extended inspection policy. *European Journal of Operational Research*, 196(3): 979–986, 2009.
6. T. Jin, F. Belkhouche, and C. H. Sung. A consolidated approach to minimize semiconductor production loss due to unscheduled ATE downtime. *Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering*, pages 188–193, sep. 2007.
7. R.C. Leachman, D. Shengwei, and C. Chen-Fu. Economic efficiency analysis of wafer fabrication. *IEEE Transactions on Automation Science and Engineering*, 4(4): 501–512, oct. 2007.
8. N. Li, M. T. Zhang, S. Deng, Z. H. Lee, L. Zhang, and L. Zheng. Single-station performance evaluation and improvement in semiconductor manufacturing: A graphical approach. *International Journal of Production Economics*, 107(2): 397–403, 2007.
9. S. Mittal and P. McNally. Line defect control to maximize yields. *Intel Technology Journal*, 4(2), 1998.
10. G.E. Moore. Progress in digital integrated electronics. *IEEE International Electron Devices Meeting*, 21: 11–13, 1975.
11. D. Ouelhadj and S. Petrovic. A survey of dynamic scheduling in manufacturing systems. Journal of Scheduling, 12: 417–431, 2009.
12. A.K. Schoemig. On the corrupting influence of variability in semiconductor manufacturing. *Proceedings of the Winter Simulation Conference Proceedings*, 1: 837–842, 1999.
13. I. Tirkel, N. Reshef, and G. Rabinowitz. In-line inspection impact on cycle time and yield. *IEEE Transactions on Semiconductor Manufacturing*, 22(4): 491–498, nov. 2009.
14. H. Wang. A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research*, 139(3): 469–489, 2002.

II.6 Supply Chain Management & Inventory

Chair: Prof. Dr. Herbert Meyr (TU Darmstadt)

This topic focuses on quantitative support of decision problems arising in the management of inter-organizational supply chains and in the long-term, strategic to short-term, operational planning of intra-organizational supply chains.

Thus, a very broad range of planning problems spanning from cooperations between independent companies to optimization and coordination of different functional departments –like procurement, production, transport and sales– within a single company is covered. Varying "production" speeds on the different stages of inter- and intra-organisational supply chains are usually balanced by inventories at stocking points. Thus, inventory management models and methods are of further special interest for this topic.

Shelf and Inventory Management with Space-Elastic Demand

Alexander H. Hübner and Heinrich Kuhn

Abstract Managing limited shelf space is a core decision in retail as increasing product variety is in conflict with limited shelf space and operational replenishment costs. In addition to their classical supply function, shelf inventories have a demand generating function, as more facings lead to growing consumer demand. An efficient decision support model therefore needs to reflect space-elastic demand and logistical components of shelf replenishment. However shelf space management models have up to now assumed that replenishment systems are efficient and that replenishment costs are not decision relevant. But shelf space and inventory management are interdependent, e.g., low space allocation requires frequent restocking. We analyzed a multi-product shelf space and inventory management problem that integrates facing-dependent inventory holding and replenishment costs. Our numerical examples illustrate the benefits of an integrated decision model.

1 Introduction

Retailers need to manage complexity to match consumer demand with shelf supply by determining the interdependent problems of assortment size, shelf space assignments and shelf replenishments. Retail shelf space assignment has demand and inventory effects. The more space is assigned to products, the higher the demand (=space-elastic demand), and potentially the lower the replenishment frequency and vice versa. Traditional shelf space management models focus on space assignment and assume efficient inventory management systems. In other words, they decouple

Alexander H. Hübner
Catholic University Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany, e-mail: alexander.huebner@ku-eichstaett.de

Heinrich Kuhn
Catholic University Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany, e-mail: heinrich.kuhn@ku-eichstaett.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_64,

the shelf space assignment decision from replenishment. We therefore propose an extension to include restocking aspects in shelf space management. Our model ensures efficient and feasible shelf inventories, clarifies restocking requirements and enables the resolution of retail-specific problem sizes.

The remainder is organized as follows. We first provide a literature review, and then develop the model in section 3. Section 4 presents computational tests. Finally, section 5 discusses potential future areas of research.

2 Literature Review

Urban [9] takes into account inventory elastic demand, since sales before replenishment reduce the number of items displayed. A sequential model first optimizes shelf space assignment, and then restocking time. Hariga et al [3] determine assortment, replenishment, positioning and shelf space allocation under shelf and storage constraints for a limited four-item case. Abbott and Palekar [1] formulate an economic order quantity problem. The optimal replenishment time has an inverse relationship with initial space assignment and space elasticity. However, it requires an initial space assignment as input, and omits inventory holding costs. Yücel et al [10] analyze an assortment and inventory problem under consumer-driven demand substitution. They conclude that neglecting consumer substitution and space limitations has significant impact on the efficiency of assortments. Hübner and Kuhn [4] extend shelf space management models with out-of-assortment substitution and ensure appropriate service levels under limited replenishment capacities.

All models assume efficient restocking and omit restocking capacity constraints. They model instantaneous and individual restocking, i.e., the retailer immediately refills the empty shelf as soon as an item runs out of stock. A detailed discussion of shelf space models can be found in Hübner and Kuhn [6].

3 Model Formulation

The majority of consumers decide on their final purchases in store and make choices very quickly after minimal searches [2]. The demand for an item i is therefore a composite function of basic demand α_i, space elasticity β_i and out-of-assortment substitution μ_{ji} from j to i.

Items can be replaced immediately from backroom stock. Showroom inventory involves a set of products $i = 1, \ldots, I$ with display capacity S. Items are moved forward to the front row to avoid partial shelf-front depletion. Common retailer practice is to conduct joint replenishment of products, e.g., in the morning before the store opens [8, 7]. The costs of joint replenishment are assumed to be independent from the facing decision and non-decision relevant. To avoid lost sales, sales employees also refill any shelf gaps that arise between the basic refilling cycles. Accordingly,

either demand d_{ik} of item i at facing level k exceeds supply q_{ik}, or vice versa. If demand exceeds supply, the insufficient basic supply requires refilling between periods t, or, if the opposite situation occurs, overstocked items increase capital employed. The third possibility, where demand matches supply exactly, is only theoretically possible, as supply is based on entire shelf slots. The following figure illustrates the development of shelf inventory levels and the associated refilling processes according to the demand-supply relationship. r_{ik} is the refilling volume and s_{ik} the overstocked volume.

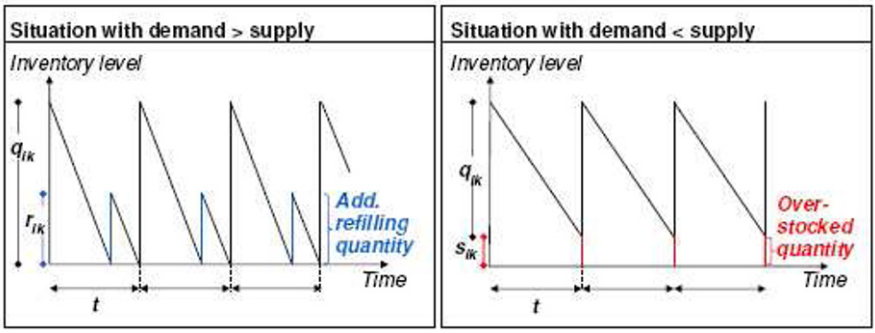


Fig. 1 Development of retail shelf inventory levels

The objective is to maximize product category profit. The *total profit* TP consists of TDP (total direct profit), TSP (total substitution profit), TCL (total listing costs), TCUS (total costs of undersupply) and TCOI (total costs of overstocked inventory).

$$\text{Max! TP} = \text{TDP} + \text{TSP} - \text{TCL} - \text{TCUS} - \text{TCOI} \tag{1}$$

The *total direct profit* covers the profit of items regardless of their relationship to the remaining assortment. d_{ik} is used to precalculate the demand for each item i and its associated facing level k, with $k = 0, 1, \ldots, K$. The preprocessing enables transfer of the non-linear demand function into a $0/1$ multi-choice knapsack problem where the binary variable y_{ik} selects the most profitable item-facing combination. p_i is the profit, b_i the breadth, and f_i the facing of the item i.

$$\text{TDP} = \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ik} \cdot d_{ik} \cdot p_i \qquad \text{with} \quad d_{ik} = \alpha_i \cdot (b_i \cdot f_i)^{\beta_i} \tag{2}$$

The *total substitution profit* integrates profit from demand shifts for delisted items. The term $(\lambda_j \cdot d_{j1})$ symbolizes the latent demand for delisted items, with λ_j expressing a share, and d_{j1} the demand at one facing. The substitution matrix μ_{ji} integrates transition probabilities between items j and i. The binary variables y_{i0} and y_{j0} (i.e., $k = 0$) indicate whether an item is listed (set to 0) or delisted (set to 1).

$$\mathrm{TSP} = \sum_{i=1}^{I} \sum_{\substack{j=1 \\ j \neq i}}^{I} (\lambda_j \cdot d_{j1}) \cdot y_{j0} \cdot \mu_{ji} \cdot (1 - y_{i0}) \cdot p_i \tag{3}$$

Product listings induce fixed costs l_i for advertisement and layout changes.

$$\mathrm{TCL} = \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ik} \cdot l_i \tag{4}$$

The *total costs of undersupply* integrate the additional refilling requirements if demand is higher than supply, expressed by the refilling volume r_{ik}. The parameter h_i describes the capacity in units behind one facing, and RFC depicts refilling costs for one shelf slot.

$$\mathrm{TCUS} = \sum_{i=1}^{I} \sum_{k=0}^{K} \frac{r_{ik}}{h_i} \cdot \mathrm{RFC} \tag{5}$$

The *total costs of overstocked inventory* comprise capital costs for overstocked volume s_{ik}, i.e., where supply exceeds demand before the next basic shelf filling process. IR is an interest rate, and c_i stands for the product costs.

$$\mathrm{TCOI} = \sum_{i=1}^{I} \sum_{k=1}^{K} s_{ik} \cdot c_i \cdot \mathrm{IR} \tag{6}$$

The constraints are composed of hierarchical planning aspects and modeling requirements. (7) ensures that only the available space S can be distributed, with b_{ik} describing the facing-dependent breadth of item i. (8) and (9) define the volumes either for under- or oversupplied volumes. (10) allows only one facing level for each item.

$$\sum_{i=1}^{I} \sum_{k=1}^{K} y_{ik} \cdot b_{ik} \leq S \tag{7}$$

$$r_{ik} \geq y_{ik} \cdot (d_{ik} - q_{ik}) + \sum_{\substack{j=1 \\ j \neq i}}^{I} \lambda_j \cdot d_{j1} \cdot y_{j0} \cdot \mu_{ji} \quad \forall i = 1,2,\ldots,I \quad \forall k = 0,1,\ldots,K \tag{8}$$

$$s_{ik} \geq y_{ik} \cdot (q_{ik} - d_{ik}) - \sum_{\substack{j=1 \\ j \neq i}}^{I} \lambda_j \cdot d_{j1} \cdot y_{j0} \cdot \mu_{ji} \quad \forall i = 1,2,\ldots,I \quad \forall k = 0,1,\ldots,K \tag{9}$$

$$\sum_{k=0}^{K} y_{ik} = 1 \qquad \forall i = 1,2,\ldots,I \tag{10}$$

$$y_{ik} \in \{0;1\} \qquad \forall i = 1,2\ldots,I \quad \forall k = 0,1,\ldots,K \tag{11}$$

$$s_{ik}, r_{ik} \geq 0 \qquad \forall i = 1,2\ldots,I \quad \forall k = 0,1,\ldots,K \tag{12}$$

4 Numerical Examples

We apply test cases to assess the performance of our integrated model. We use data with a high profit-space correlation to evaluate the performance of a hard knapsack problem as in Hübner and Kuhn [5].

The test cases with 25 items demonstrate the profit impact over current industry practice, and the value of an integrated restocking and shelf space assignment model. Total profit increases by 17.5% compared to industry practice (lower bound), which could be quite substantial in low-margin industry retailing. Secondly, the example shows the value of an integrated restocking and shelf space assignment model. The model LS optimizes only for listing and spacing, i.e., supply costs are not part of the objective function, and are calculated a posteriori. Here, the test case reveals high undersupply costs leading to lower total profit. Disregarding supply costs results in frequent restocking needs.

Our tests also show that the solution structure (i.e., facing levels) changes significantly. 56% of the items receive different facing levels in the optimized LSR model compared to the lower bound.

Table 1 Numerical examples

Model [a]	**Lower bound**	**LS model**	**LSR model**
TDP [EUR]	343,275	404,177	408,669
TSP [EUR]	12,218	12,524	3,573
TCL [EUR]	2,600	3,400	4,000
TCUS [EUR]	4,850	20,396	10,159
TCOI [EUR]	9,376	11,731	162
Total [EUR]	338,667	381,175	397,921

[a] Lower bound: "space to sales" logic as in commercial software
LS: listing and spacing; LSR: listing, spacing and restocking

5 Conclusions and Future Research

Our model extends known shelf space models with replenishment costs and clarifies restocking requirements. It has been solved with CPLEX, and allows the computation of optimal results for category-specific problem sizes. The numerical example shows the benefits of an integrated model over current industry practice and aligns space and restocking requirements.

Areas of further research lie in investigation of the joint optimization of space assignment, instore replenishment cycles and order cycles for backroom replenishment. Additional possibilities are the integration of backroom capacities and inventory costs. Competitive scenarios and demand-influencing marketing effects could be part of an integrated analysis, as well as an extension to stochastic demand models.

References

1. Harish Abbott and Udatta S. Palekar. Retail replenishment models with display-space elastic demand. *European Journal of Operational Research*, 186(2): 586–607, 2008.
2. Pierre Chandon, Wesley J. Hutchinson, Eric T. Bradlow, and Scott H. Young. Does in-store marketing work? Effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase. *Journal of Marketing*, 73(November): 1–17, 2009.
3. Moncer A. Hariga, Abdulrahman Al-Ahmari, and Abdel-Rahman A. Mohamed. A joint optimisation model for inventory replenishment, product assortment, shelf space and display area allocation decisions. *European Journal of Operational Research*, 181(1): 239–251, 2007.
4. Alexander H. Hübner and Heinrich Kuhn. Retail shelf space management model with integrated inventory-elastic demand and consumer-driven substitution effects. *http://ssrn.com/abstract=1534665*, 2009.
5. Alexander H. Hübner and Heinrich Kuhn. Integrated retail shelf space and price management. *Technical Report KU Eichstätt-Ingolstadt*, pages 1–27, 2010.
6. Alexander H. Hübner and Heinrich Kuhn. Quantitative models for retail category management: A review of assortment and shelf space planning in practice, software applications and science. *http://ssrn.com/abstract=1579911*, 2010.
7. Gürhan A. Kök and Marshall L. Fisher. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6): 1001–1021, 2007.
8. Stephen A. Smith and Narendra Agrawal. Management of multi-item retail inventory systems with demand substitution. *Operations Research*, 48: 50–64, 2000.
9. Timothy L. Urban. An inventory-theoretic approach to product assortment and shelf-space allocation. *Journal of Retailing*, 74(1): 15–35, 1998.
10. Eda Yücel, Fikri Karaesmen, F. Sibel Salman, and Metin Türkay. Optimizing product assortment under customer-driven demand substitution. *European Journal of Operational Research*, 199(3): 759–768, 2009.

Quantity Flexibility for Multiple Products in a Decentralized Supply Chain

İsmail Serdar Bakal and Selçuk Karakaya

1 Introduction

One of the major complicating factors in decentralized supply chains is the long procurement/manufacturing lead times, which forces the upstream members to commit resources to production based on forecasted demand. Downstream members (the retailer for the rest of the manuscript) would like to have a higher production quantity to be able to satisfy the demand whereas upstream members (the manufacturer for the rest of the manuscript) would like to have some sort of assurance about the demand so that they will not be building unnecessary capacity. Traditionally, such a conflict is resolved by an initial estimate provided by the retailer. However, the manufacturer is aware that the retailer is likely to manipulate this initial order; hence, the initial estimate provides little incentive for the manufacturer to build the capacity that the retailer would like to have. One resolution offered in the literature to overcome this issue is the Quantity Flexibility contract, where the retailer guarantees to order no less than a certain percentage of the initial estimate and the manufacturer guarantees to deliver a certain percentage above.

Quantity Flexibility contracts and their extensions have been widely investigated in the literature. Some examples are [2], [6], [5], [4], [7], [1] and [3]. Note that all these studies consider a single-item system whereas our aim is to investigate the characteristics of a quantity flexibility contract involving two items that share common features. Our study is motivated by the interaction between an automobile manufacturer and the sales headquarters. Sales headquarters place initial orders for different models which share a significant proportion of common parts. Towards the selling season, the headquarters have the opportunity to revise their orders to

İsmail Serdar Bakal
Middle East Technical University, ODTU Endustri Muhendisligi Bolumu 06531 Ankara, Turkey, e-mail: bakal@ie.metu.edu.tr

Selçuk Karakaya
Middle East Technical University, ODTU Endustri Muhendisligi Bolumu 06531 Ankara, Turkey, e-mail: karakaya.selcuk@gmail.com

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 411
DOI 10.1007/978-3-642-20009-0_65,

some extent with the constraint that final aggregate order should match the initial aggregate order.

In this study, we present a stylized model for such an environment and characterize the optimal order quantities for the retailer and manufacturer. We also analyze the benefits of such a flexibility scheme by comparing it to a no-flexibility setting where the retailer orders in advance.

2 Problem Description and Assumptions

We analyze a decentralized supply chain with a single retailer and a single manufacturer where the retailer sells two products to an end market in a single period. The retailer firstly quotes an initial order for each product based on the demand forecast. The sum of initial orders determines the total final order quantity for the selling period. Given the initial order, the manufacturer begins to install its production capacity and procure components by regular delivery mode for production. Meanwhile, the retailer collects more market demand information before committing the final order of each product. In the second decision stage, although the retailer is not allowed to change the total quantity of the order; he can adjust the order quantity of each product freely as long as the sum of final orders is equal to the sum of initial orders. The manufacturer is obligated to fill the retailer's final order for each product. She utilizes expedited delivery if required, to procure additional components. The expedited delivery provides shorter procurement lead time but its cost is higher than regular delivery. Main characteristics and assumptions of the model are:

- The demand of each product is assumed to be independent of the demand of the other product. The retailer acquires perfect demand information before submitting final order.
- The retailer's cost includes only the unit purchase cost. The retailer's revenue includes the unit selling price from the products sold during the selling season. At the end of the selling season, if there are any items in inventory, they are cleared out at a discounted price.
- The manufacturer has two options to procure components of products. The first one is regular delivery at the beginning of the season. The other one is expedited delivery after the retailer places the final orders. The expedited delivery is more expensive than regular delivery.
- The sum of final orders of the retailer cannot exceed the sum of the initial orders even if it is beneficial for both parties.
- The manufacturer's cost includes unit production cost and unit procurement cost for regular and expedited delivery. The manufacturer's revenue includes the unit wholesale price from the products sold to the retailer. Any unsold item will be cleared out at a discounted price (Please refer to Table 1 for a summary of the parameters and decision variables).

Table 1 Summary of parameters and decision variables

p_i	Unit price of product i
s_i	Discounted price of product i
w_i	Wholesale price of product i
m_i	Manufacturer's unit procurement cost for product i
r_i	Manufacturer's discounted sales price for product i at the end of the period
d_1	Cost of regular delivery
d_2	Cost of expedited delivery
X_i	Random variable denoting the demand of product i at the beginning of the period with pdf $f_i(x)$ and cdf $F_i(x)$.
x_i	The realization of X_i
Q	Sum of retailer's initial orders
q_i^r	Retailers final order quantity for product i
q_{1i}^m	Manufacturer's regular order for product i
q_{2i}^m	Manufacturer's expedited order for product i

The cost and revenue parameters are assumed to be as $p_i > w_i > s_i$ for the retailer, and similarly as $w_i > m_i + d_1 > r_i$ for the manufacturer. The delivery cost parameters are assumed to be $d_2 > d_1$. The sequence of events is illustrated in Figure 1.

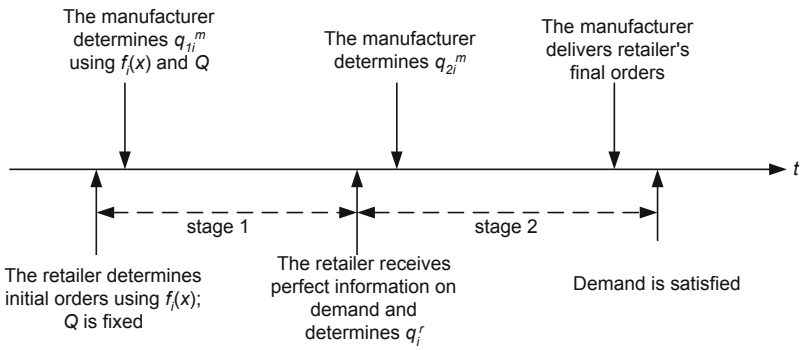


Fig. 1 Sequence of Events

3 Analysis of Order Adjustment Flexibility

Note that the only restriction on the final order quantities of the retailer is the sum of initial orders. Hence, the relevant decision variable for the retailer is the sum of initial orders, Q, rather than the individual orders. Since its final orders should sum up to Q as well, the retailer may end up with excess inventory, even with perfect

demand information. Such a case would occur when the sum of demand realizations turns out to be less than Q. Without loss of generality, we assume that product 1 is more profitable during regular sales whereas product 2 is more profitable at the discounted sales. That is, $p_1 - w_1 > p_2 - w_2$, and $w_1 - s_1 > w_2 - s_2$. We begin our analysis with the retailer's final order quantities, given the aggregate order quantity specified in the first stage, Q, and the demand realization, x_1 and x_2:

$$(q_1^{r*}, q_2^{r*}) = \begin{cases} (Q,0) & x_1 \geq Q \\ (x_1, Q - x_1) & x_1 < Q \end{cases}. \tag{1}$$

We now consider the expedited order quantities of the manufacturer. Note that the manufacturer is obligated to fill the retailer's final order for each product. As a result, given the manufacturer's initial procurement quantity, the manufacturer will set the second procurement level to meet the retailer's final order quantities.

(i) If $q_{11}^m + q_{12}^m \geq Q$

$$(q_{21}^{m*}, q_{22}^{m*}) = \begin{cases} (Q - q_{11}^m, 0) & x_1 > Q \\ (x_1 - q_{11}^m, 0) & q_{11}^m < x_1 < Q \\ (0,0) & Q - q_{12}^m < x_1 < q_{11}^m \\ (0, Q - x_1 - q_{12}^m) & x_1 < Q - q_{12}^m \end{cases} \tag{2}$$

(ii)If $q_{11}^m + q_{12}^m < Q$

$$(q_{21}^{m*}, q_{22}^{m*}) = \begin{cases} (Q - q_{11}^m, 0) & x_1 > Q \\ (x_1 - q_{11}^m, 0) & Q - q_{12}^m < x_1 < Q \\ (x_1 - q_{11}^m, Q - x_1 - q_{12}^m) & q_{11}^m < x_1 < Q - q_{12}^m \\ (0, Q - x_1 - q_{12}^m) & x_1 < Q - q_{11}^m \end{cases} \tag{3}$$

We can now incorporate the optimal solutions of the second stage problems to the first stage problem to determine the optimal initial orders. The expected profit of the retailer in the first stage is given by

$$\Pi_r(Q) = \int_0^Q \left[(p_1 - w_1 + w_2)x_1 - w_2 Q + \int_0^{Q - x_1} [p_2 x_2 + s_2(Q - x_1 - x_2)] f_2(x_2) dx_2 \right.$$
$$\left. + \int_{Q - x_1}^{\infty} p_2 (Q - x_1) f_2(x_2) dx_2 \right] f_1(x_1) dx_1 + \int_Q^{\infty} (p_1 - w_1) Q f_1(x_1) dx_1$$

Proposition 1 *The optimal initial order quantity of the retailer, Q, is characterized by the unique solution of $d\Pi_r(Q)/dQ = 0$.*

Proof.

$$\frac{d\Pi_r(Q)}{dQ} = p_1 - w_1 - (p_1 - w_1 - p_2 + w_2)F_1(Q) - \int_0^Q (p_2 - s_2)F_2(Q - x_1)f_1(x_1)dx_1$$

$$\frac{d^2\Pi_r(Q)}{dQ^2} = -(p_1 - w_1 - p_2 + w_2)f_1(Q) - \int_0^Q (p_2 - s_2)f_2(Q - x_1)f_1(x_1)dx_1 < 0$$

Since $\lim_{Q\to 0} d\Pi_r(Q)/dQ = p_1 - w_1 > 0$ and $\lim_{Q\to\infty} d\Pi_r(Q)/dQ = -w_2 + s_2 < 0$, the solution to $d\Pi_r(Q)/dQ = 0$ is unique and optimal. □

From the first order condition, we observe that the retailer's order quantity increases in the regular sales prices, and decreases in the transfer prices. The discounted sales price of product 1 does not affect it whereas the discounted sales price of product 2 has an increasing effect.

The expected profit of the manufacturer in the first stage is given by

$$\begin{aligned}\Pi_m(q_{11}^m, q_{12}^m) = & -(m_1 + d_1)q_{11}^m - (m_2 + d_1)q_{12}^m + \int_0^Q [w_1 x_1 + w_2(Q - x_1)]dF_1(x_1) \\ & + \int_Q^\infty [w_1 Q - (m_1 + d_2)(Q - q_{11}^m) + r_2 q_{12}^m]dF_1(x_1) + \int_0^{q_{11}^m} r_1(q_{11}^m - x_1)dF_1(x_1) \\ & + \int_{Q - q_{12}^m}^Q r_2(q_{12}^m - Q + x_1)dF_1(x_1) - \int_{q_{11}^m}^Q (m_1 + d_2)(x_1 - q_{11}^m)dF_1(x_1) \\ & - \int_0^{Q - q_{12}^m} (m_2 + d_2)(Q - x_1 - q_{12}^m)dF_1(x_1) \end{aligned} \tag{4}$$

Proposition 2 *Let* $(q_{11}^m)'$ $(q_{12}^m)'$ *be the unique solution to* $\partial\Pi_m/\partial q_{11}^m$ *and* $\partial\Pi_m/\partial q_{12}^m$. *Then, the optimal initial procurement quantities of the manufacturer are* $(q_{11}^m)^* = \min(Q, (q_{11}^m)')$ *and* $(q_{12}^m)^* = (q_{12}^m)'$.

Proof. The second order conditions indicate that $\Pi_m(q_{11}^m, q_{12}^m)$ is jointly concave. Note that optimal order quantities of the manufacturer will not be greater than Q because of the retailer's final order quantities. If $(q_{11}^m)' \le Q$, first order conditions provide the optimal order quantities. Otherwise, we will have $(q_{11}^m)^* = Q$. □

The optimal initial order quantities of the manufacturer do not depend explicitly on w_1 and w_2 since the sales quantity is directly related to the retailer's final orders. However, it should be noted that they depend on Q, which depends on both w_1 and w_2. It should be also noted that they depend only on the distribution of the demand of product 1. This results from the fact that the final order quantities of the

retailer are determined by the demand of product 1. That is, if is sufficient to cover the demand for product 1, the remaining portion of the order is filled by product 2 regardless of its demand realization. The effects of an increase in each of the various parameters on the optimal order quantities of the manufacturer are summarized as follows:

	d_1	d_2	m_1	m_2	r_1	r_2	w_1	w_2	s_1	s_2	p_1	p_2
q^m_{11}	↘	↗	↘	→	↗	→	↘	→	↘	→	→	↗	→	↗	→	↗	→
q^m_{12}	↘	↗	→	↘	→	↗	↘		↘		→	↗		↗		↗	

4 Conclusion

In this study, we investigated a quantity flexibility contract between a retailer and a manufacturer that involves multiple products sharing common parts. We analyzed the resulting model and characterized the optimal order quantities of both parties. Effects of various parameters on the order quantities of the retailer and manufacturer are described analytically. We also quantified the magnitude of the improvement in expected profits due to flexibility by comparing the quantity flexibility scheme to a no-flexibility setting through a computational study (the details of the computational study is excluded due to space limitations). Our findings indicate that the quantity flexibility contract always benefits the retailer as expected. Although the effects of quantity flexibility on the manufacturer's profit depend on the problem parameters, there are numerous settings where quantity flexibility increases her profits significantly.

References

1. Y. Bassok and R. Anupindi. Analysis of Supply Contracts with Commitments and Flexibility. *Naval Research Logistics*, 55: 459–477, 2008.
2. Y. Bassok, A. Bixby, R. Srinivasan, and H.Z. Wiesel. Design of Component-Supply Contract with Commitment-Revision Flexibility. *IBM Journal of Research and Development*, 41(6): 693–703, November 1997.
3. Z. Lian and A. Deshmukh. Analysis of Supply Contracts with Quantity Flexibility. *European Journal of Operational Research*, 196: 526–533, 2009.
4. J.M. Milner and P. Kouvelis. Order Quantity and Timing Flexibility in Supply Chains: The Role of Demand Characteristics. *Management Science*, 51(6): 970–985, June 2005.
5. S.P. Sethi, H. Yan, and H. Zhang. Quantity Flexibility Contracts: Optimal Decisions with Information Updates. *Decision Sciences*, 35(4): 691–712, Fall 2004.
6. A. A. Tsay. The Quantity Flexibility Contract and Supplier-Customer Incentives. *Management Science*, 45(10): 1339–1358, October 1999.
7. J. Wu. Quantity Flexibility Contracts under Bayesian Updating. *Computers and Operations Research*, 32: 1267–1288, 2005.

Achieving Better Utilization of Semiconductor Supply Chain Resources by Using an Appropriate Capacity Modeling Approach on the Example of Infineon Technologies AG

Hans Ehm, Christian Schiller, and Thomas Ponsignon

Abstract The capacity model is a crucial point for the long-term resource planning and the demand-supply match. So far, Infineon has been using a modeling approach based on aggregated capacity groups (ACG), which encompass products that are interchangeable in production resources. Using ACGs provides flexibility since demands can be switched within the group without altering the supply plan. However, this model lacks visibility when the product mix is changing. In this paper we introduce a different capacity modeling concept that makes all critical resources visible. The model has been successfully implemented within Infineon planning landscape, and its performance has been tested for backend resources. Results show that the new approach achieves 6% more production load per week with 14% less capacity. Knowing the level of investments for production capacities in the semiconductor industry, this improvement has a significant financial impact for the company.

1 Introduction

This work is motivated by the difficulty of finding an appropriate capacity modeling approach for allocating supply chain resources. It is even more arduous for complex manufacturing systems such as semiconductor supply chains. Four major phases are needed to obtain semiconductor devices from raw silicon wafers. It starts with the wafer fabrication (Fab) during which their surface is modified to create patterns of integrated circuits. Afterwards the individual dies on the wafers are tested (Sort). The frontend operations are completed after this step. The dies are cut and stored in a die-bank that serves as a buffer stock before backend operations. During the assembly process (Assy) the dies are encapsulated into synthetic packages to pro-

Hans Ehm, Christian Schiller, Thomas Ponsignon
Infineon Technologies AG, Am Campeon, 85579 Neubiberg, Germany, e-mail: hans.ehm@infineon.com, christian.schiller@infineon.com, thomas.ponsignon@infineon.com

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 417
DOI 10.1007/978-3-642-20009-0_66,

tect them from environmental influences. Finally the finished chips are tested for functionality (Test) and shipped to distribution centers and to customers [10].

The planning of resources in semiconductor supply chain is challenging because of several factors. First, it usually involves dozens of in-house frontend and backend facilities plus silicon foundries and subcontractors spread all over the world, which are organized as a global network. Then, it has to be dealt with the long lead times (up to four months), while products tend to have always shorter life cycles and steeper production ramp-ups. In addition, semiconductors are certainly one of the most volatile markets with cyclic up- and downturn phases, which make the demand very difficult to predict. Furthermore, capacity expansions are typically long and expensive, i.e. up to $18 million for certain frontend tools and $4 billion for a wafer fab. As a consequence, an appropriate utilization of the resources of the supply chain is one of the key success factors in the semiconductor industry [10, 5, 3].

There are some previous works related to resource planning in semiconductor environment. We distinguish between publications that focus on long-term strategic decisions and other papers that consider the mid-term operational demand-supply matching. Bermon and Hood [2] propose a linear program for optimizing strategic capacity allocation at an IBM wafer fab. This work has been extended by integrating stochastic demand in Barahona et al [1]. Denton et al [4] suggest a mixed-integer programming formulation and corresponding heuristics for both strategic and operational levels in semiconductor supply chains. The model contains details, such as discrete lot sizes, for daily decisions. Habla and Mönch [6] introduce a linear program that allows solving volume and capacity planning problems. They use a general product structure and they investigate the required level of detail for bottleneck modeling. Finally, Kallrath and Maindl [7] study the use of the SAP$\ddot{o}$ APO tool for production planning in semiconductor environment. Their analysis reveals that it is solved by a purely rule-based algorithm called Capable-to-Match.

2 Actual Capacity Modeling Approach at Infineon

2.1 Infineon's Supply Chain Planning Processes

The focus of this paper is on planning activities. The purpose of supply chain planning is to fulfill customer demands through efficient utilization of the resources of the company and of its partners. According to the Supply Chain Operations Reference (SCOR) framework model by the Supply Chain Council [9], the planning activities can be divided into five sub-processes: (1) long-term demand planning that deals with aggregated supply chain requirements, (2) long-term and mid-term capacity provision for aggregated supply chain resources, (3) mid-term demand-supply match, which balances supply chain resources with requirements on a more detailed level, (4) short-term order management that communicates supply chain

plans towards customers, and (5) short-term production management, which provides load plans to manufacturing sites.

The capacity model is a crucial point for the second and the third sub-processes as it states the main quantitative restriction of the resource optimization problem. Indeed, tactical decisions (e.g. investment for machine purchasing, share between in-house production and outsourcing, headcount planning) as well as operational decisions (e.g. balancing of weekly production load, securing minimum stocks and deliveries) are taken during the resource provision process based on aggregated capacity corridors. The latter are first communicated by each facility and then consolidated on corporate level. In addition, capacities play a major role for the demand-supply alignment since customer orders and forecasts are confirmed or postponed in concordance with these bounds [8]. We will now describe more precisely the capacity model that is used nowadays at Infineon.

2.2 *Current Capacity Model at Infineon: ACG*

So far Infineon has been using a capacity modeling approach based on Aggregated Capacity Groups (ACGs). An ACG encompasses products that are interchangeable in term of production resources as they use similar equipment with identical capacity consumption. Along the manufacturing process there are only two stages where ACGs can be modeled, namely Fab and Assy. Hence, a Fab-ACG is defined as a group of identical wafer processes, whereas an Assy-ACG is a group of similar chip packages. There is no dedicated ACG for Sort or Test. Their resources are simply considered as infinite.

Fig. 1 shows an example of two Assy-ACGs that use the same equipment with their respective product assignment and capacity consumption. The upper limit of each ACG is based on the capacity feedback provided by the production sites during the capacity planning process by using a rough demand estimation on a highly aggregated level.

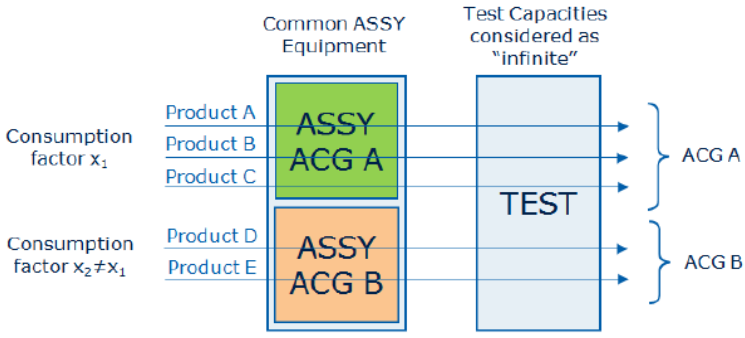


Fig. 1 Example of two Assy-ACGs.

2.3 Advantages and Disadvantages of ACGs

Using an ACG as higher aggregation level instead of single item granularity provides more flexibility for supply chain planning since demands for these products can be switched within the group without altering the supply plan.

However, ACGs also show some practical drawbacks that moderate their theoretical flexibility. Indeed, the fact that ACGs can only be modeled at some selected points of the manufacturing process leads to an unrealistic capacity model since bottlenecks may actually occur at every stage. Ideally, the number of bottlenecks should not be limited at all so that one stage can have more than one capacity restriction, e.g. both lithography and furnace in Fab-process.

Since ACGs are set up during the capacity planning process that is only dealing with forecasts and stock targets, the decisions on capacity groups are taken without regard to real market demand and priorities. In fact, customer orders and reservations are primary considered one step later during the demand-supply match. As a consequence, forecast bias may lead to inappropriate decisions on capacity groups that result in unrealistic supply plans. Furthermore, additional production constraints are not considered during the setting of ACGs, e.g. special material requirements for low volume products, environmental criterion that distinguishes between "gray" and "green" chips with regard to the level of lead in the plating. These lacks decrease the consistency and the quality of the generated supply plan.

Moreover, the mapping of products in ACGs is directly implemented in the databases. But with several thousand of planning items and an average product life cycle of 18 months, it becomes very laborious to maintain the groups.

Finally, ACGs are seen as rigid structures that split the available resources of the company into limited capacity buckets. They hinder from optimizing the demand-supply match globally. With this, it becomes clear that a new approach is required, which is introduced in next section.

3 Proposed Alternative to ACGs

3.1 Description of the New Approach

As a first step, we calculate the capacity of each workstation along the manufacturing process (e.g. lithography, bumping, moldingĚ). For this, we multiply the total number of lines or machines that are available within the network by the Overall Equipment Efficiency (OEE), which states the time during which products can be processed. Then, from the production recipes we derive the consumption factors (CF) of the products. The latter are used for converting the available machine-time into maximum throughput in pieces per period.

Based on the product-mix provided by the weekly demand-supply match process, the loading level of each workstation is assessed. Each stage that exceeds 85% of

utilization is considered as a bottleneck resource, and it is explicitly modeled as such in the system. All other machines are seen as non-critical tools, and they are just ignored. Note that the threshold is an empirical value that can be slightly adjusted depending on the complexity of the product-mix. After this step, the system used for the demand-supply match comprises throughputs and consumption factors for all critical resources.

Afterwards, we deal with prioritized demand that is composed of customer orders, reservations, additional forecasts, and stock replenishments. This step corresponds to the demand-supply match. It is executed by means of a step-by-step heuristic procedure based on rules that tend to maximize capacity utilization and to minimize unfulfilled demand. Finally, the resulting supply plan is provided to factories and to the order management system.

Fig. 2 shows an example of two Assy-bottlenecks and one Test-Bottleneck with the respective consumption factors of the products. The new approach allows calculating more realistic production plans since it is no longer based on rough outcomes of the capacity planning process but on real customer demands.

3.2 Proof of Concept and Benefits

The new model does not create artificial capacity constraints that hinder demands to be planned. Moreover, all critical resources can be made visible instead of lumped together, and they are constraining the production requests for a more realistic supply plan. The new capacity representation is more complex to track as a complete demand-supply match process takes place every week, but it is supported by systems providing better results than human decisions.

To confirm the advantages of the new approach, a pilot project has been conducted with real-world figures at Infineon for the production planning in backend of sensor products. The scope of the analysis is restricted to 87 final items that can be

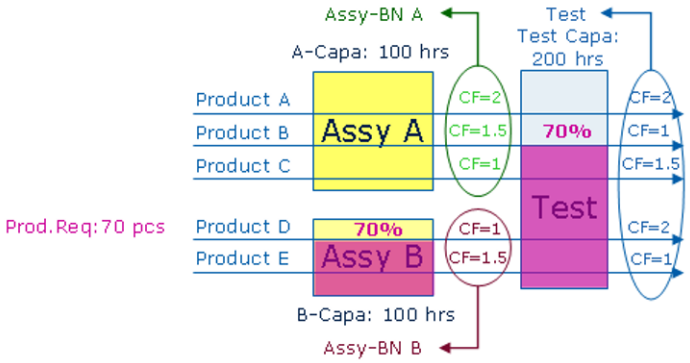


Fig. 2 The new approach shows the impact of each production request on every bottleneck load.

processed on five assembly lines and ten testers. On the one hand, the products are sorted, according to the former capacity model, in two Assy-ACGs with regard to their consumption factors. On the other hand, the procedure described in Sect. 3.1 is executed aside in the tool that is used for the demand-supply match. The comparison is carried out for 26 weeks. It comes out that the new approach achieves in average 6% more production load per week with 14% less capacity. Thus, the lateness of orders is drastically reduced. In addition, the new capacity model allows a 100% utilization rate, while the ACG-model can only use 81.6% of the available capacity. Knowing the level of investments for building and maintaining production capacities in the semiconductor industry, this improvement has a significant financial impact for the company.

4 Conclusion and Next Steps

We presented an alternative capacity modeling approach to the unsatisfactory concept of ACGs that Infineon has been using so far for supply chain planning activities. The new model uses an explicit representation of all bottleneck resources. Hence, it allows providing more realistic supply plans. Experiments were performed during a pilot project to assess the benefits. As next step we intent on rolling out the implementation of the new method for further product lines.

References

1. F. Barahona, S. Bermon, O. Günlük, and S. Hood. Robust Capacity Planning in Semiconductor Manufacturing. *Naval Research Logistics*, 52: 459–468, 2005.
2. S. Bermon and S. J. Hood. Capacity Optimization Planning System (CAPS). *Interfaces*, 29(5): 31–50, 1999.
3. C.-F. Chien, S. Dauzère-Pérès, H. Ehm, J. W. Fowler, Z. Jiang, S. Krishnaswamy, L. Mönch, and R. Uzsoy. Modeling and Analysis of Semiconductor Manufacturing in a Shrinking World: Challenges and Successes. *Proceedings of the Winter Simulation Conference*, pages 2093–2099, 2008.
4. B. T. Denton, J. Forrest, and R. J. Milne. IBM Solves a Mixed-Integer Program to Optimize Its Semiconductor Supply Chain. *Interfaces*, 36(5): 386–399, 2006.
5. J. N. D. Gupta, R. Ruiz, J. W. Fowler, and S. J. Mason. Operational Planning and Control of Semiconductor Wafer Fabrication. *Production Planning and Control*, 17(7): 639–647, 2006.
6. C. Habla and L. Mönch. Solving Volume and Capacity Planning Problems in Semiconductor Manufacturing: A Computational Study. *Proceedings of the Winter Simulation Conference*, pages 2260–2266, 2008.
7. J. Kallrath and T. I. Maindl. Real Optimization with SAP®APO. Springer, 2006.
8. C. Schiller. Practical Problems in Implemented Decisions Supporting Systems within the Planning Process at Infineon Technologies AG. *Intelligente Systeme zur Entscheidungsunterstützung*, pages 99–112, 2008.
9. Supply Chain Council. SCOR, 2010. http://supply-chain.org/scor.
10. R. Uzsoy, C.-Y. Lee, and L. A. Martin-Vega. A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation and Production Planning. *IIE Transactions*, 24(4): 47–60, 1992.

Towards Leveled Inventory Routing for the Automotive Industry

Martin Grunewald, Thomas Volling, and Thomas S. Spengler

Abstract Motivated by the success of Japanese car manufacturers, there is an increasing interest in the introduction of leveled logistics concepts. These concepts are characterized by leveled quantities and periodic pick up and arrival times. As a consequence, new requirements regarding the planning of material replenishment result. Planning systems based on the sequential MRP-logic do not take leveling into account and might therefore result in inacceptable volatility. To this end, we propose a modeling framework for leveled replenishment. Building blocks comprise a module for vehicle routing, where cost optimal routes are determined, and inventory control, where leveled quantities are generated. The approach extends the well-known concepts of vehicle routing with respect to the requirements of the automotive industry.

1 Introduction

The (German) automotive industry is characterized by a high product variety. This leads to a large number of different parts and variants procured from a network of suppliers. To cope with this variety, material planning systems based on the sequential MRP-logic are used. These planning systems consistently link customer orders to parts supply and do not take leveling aspects into account. As a consequence, inacceptable volatility results and thus system dynamical problems, such as the bullwhip effect, might follow. Motivated by the success of Japanese car manufacturers, there is an increasing interest in the introduction of leveled replenishment. These concepts are characterized by leveled quantities and periodic pick up and arrival times. Advantages of leveling are more stable processes and less system dynamical

Martin Grunewald · Thomas Volling · Thomas S. Spengler
Institute of Automotive Management and Industrial Production, Technische Universität Braunschweig, Katharinenstr. 3, 38106 Braunschweig, e-mail: m.grunewald | t.volling | t.spengler@tu-bs.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 423
DOI 10.1007/978-3-642-20009-0_67,

effects. However, new requirements regarding the planning of material replenishment result.

To this end, we propose a modeling framework for leveled replenishment. Building blocks comprise a module for vehicle routing, where cost optimal routes are determined, and inventory control, where leveled quantities are generated. By integrating both modules we ensure periodic pick up and arrival times with leveled quantities, while taking into account uncertainty. The approach extends the well-known concepts of vehicle routing with respect to the requirements of the automotive industry.

The paper is organized as follows. In Section 2 we describe the leveled inventory control and show the connection to vehicle routing. An integrated mathematical model is formulated in Section 3. We furthermore give an example to demonstrate the advantages of leveling. Section 4 features conclusions.

2 Leveled Inventory Control and Vehicle Routing

Inventory management is primarily concerned with finding the optimal inventory level to compensate uncertainty regarding demand, lead time and delivery quantity. Inventory control is an important instrument to achieve this objective. It determines when a replenishment order is placed and how large the order should be. In the German automotive industry orders are currently placed at fixed times and the order quantity is computed by an MRP-logic [1]. As a result, stochastic demand for product options affects the order quantity. The fluctuating order quantity propagates in the supply chain and system dynamical effects lead to inefficiencies such as overcapacities and excess stock at the suppliers.

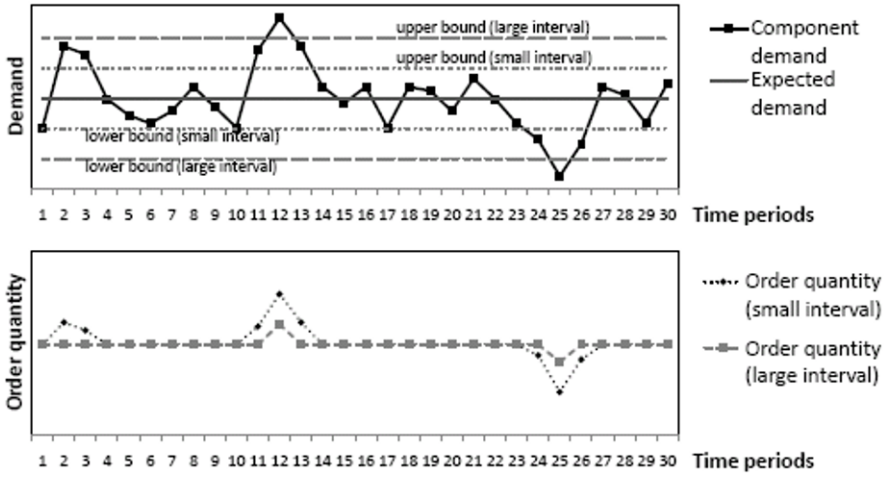


Fig. 1 Example of two control band policies with different interval widths

Against this background, leveled inventory control seeks to minimize system dynamical effects. The underlying concept is to order the same quantities at fixed periodic times. Methods that determine these fixed order quantities and periods are needed. Furthermore, rules must be defined to deal with exceptions.

Similar to Ormerci [2] we present a leveling order policy in the following. The idea is to introduce a control band. This band defines a symmetric interval around the expected demand. Following a control band policy, the fixed order quantities, which is equal to the average demand, are only adjusted, if the actual demand exceeds or falls below the thresholds defined by the band. It therefore acts as a filter. The amount by which the fixed order quantity is adjusted corresponds to the difference between the interval bound and the demand. Figure 1 shows an example for two control band policies with a small and a large interval width.

This approach is closely linked to logistics, since the order quantities are the input for vehicle routing. With increasing interval width, the variability of the order quantity is decreasing (see figure 1). A high degree of leveling is advantageous for routing since overcapacity, which is needed to compensate quantity fluctuations, can be reduced. On the downside leveling requires safety stock to hedge against demand variability. For this reason there are conflicting targets between the transportation costs which are low for a high degree of leveling and the safety stock costs which are high in that case. Therefore, there is a need for coordinated planning.

3 Integrating Leveling Order Policy into a Stochastic Vehicle Routing Problem

In this section we formulate a mathematical model to integrate a control band policy into vehicle routing with the objective to minimize transportation and safety stock costs. To simplify the model formulations we deal with single item problems. The degree of leveling and the routes are the degrees of freedom to be determined.

The basic model is the capacitated vehicle routing problem (CVRP) [4]. The objective is to minimize the transportation costs C_{trans} (see (1)) under capacity constraints (2). The optimal solution must be a valid solution of the multiple traveling salesman problem (see (3)).

$$\min_{x} \sum_{k} \sum_{i,j} c_{ij} x_{ijk}, \tag{1}$$

$$\text{s.t.} \sum_{i,j} d_i x_{ijk} \leq Q,\ k = 1, 2, \ldots, m, \tag{2}$$

$$x = \left[x_{ijk}\right] \in S_m, \text{ where} \tag{3}$$

c_{ij} = the cost of traveling from i to j,
$x_{ijk} = \begin{cases} 1, \text{ if vehicle k travels from i to j,} \\ 0, \text{ otherwise,} \end{cases}$
m = the number of the vehicles available,
S_m = the set of all feasible solutions to the m-traveling salesman problem,
d_i = the amount demanded at location i, and
Q = the vehicle capacity.

If demand is stochastic then a stochastic CVRP (SCVRP) results. Stewart and Golden present a chance-constrained model for the SCVRP. The capacity constraint (2) must be fulfilled at least with a probability of $1-\alpha$.

$$\min_x \sum_k \sum_{i,j} c_{ij} x_{ijk}$$
$$\text{s.t. } Pr\left(\sum_{i,j} D_i x_{ijk} \le Q\right) \ge 1-\alpha,\ k=1,2,\ldots,m,$$
$$x = \left[x_{ijk}\right] \in S_m,$$
where D_i is the stochastic demand at location i.

In our model the stochastic demand D_i is replaced by the order quantity Γ_i. The order quantity Γ_i depends on the stochastic demand D_i and the width a of the symmetric interval of the control band policy. Accordingly the order quantity is also a random variable. If μ_i denotes the expected value of D_i then Γ_i is determined by

$$\Gamma_i = f(D_i,a) = \begin{cases} \mu_i + \overbrace{(D_i - (\mu_i + \frac{a}{2}))}^{\text{exceedance of upper bound}} & ,\text{if } D_i \ge \mu_i + \frac{a}{2}, \\ \mu_i & ,\text{if } \mu_i - \frac{a}{2} \le D_i \le \mu_i + \frac{a}{2}, \\ \mu_i - \underbrace{((\mu_i - \frac{a}{2}) - D_i)}_{\text{shortfall of lower bound}} & ,\text{if } D_i \le \mu_i - \frac{a}{2}. \end{cases}$$

To integrate the control band policy and the interval width a, the objective function (1) is expanded by the costs for the safety stock $C_{ss} = h \cdot z \cdot \sigma_{hd}(a)$, where h is inventory costs per unit, z is the safety factor and $\sigma_{hd}(a)$ is the standard deviation of the error between order quantity and demand which is hedged by the safety stocks depending on a [3]. The standard deviation $\sigma_{hd}(a)$ rises when the degree of leveling is increased and thus also increasing the safety stock. The chance-constrained formulation for the leveled inventory routing problem is given by

$$\min_{x,\,a} \sum_k \sum_{i,j} c_{ij} x_{ijk} + h \cdot z \cdot \sigma_{hd}(a)$$
$$\text{s.t. } Pr\left(\sum_{i,j} f(D_i,a) x_{ijk} \le Q\right) \ge 1-\alpha,\ k=1,2,\ldots,m, \tag{4}$$
$$x = \left[x_{ijk}\right] \in S_m.$$

Note that the interval width a is a decision variable of the model (4). And the probability distribution of the order quantity and the safety stock costs depend on a. The following example shows the influence of a on the optimal capacity solution. Figure 2 depicts a supplier network with five suppliers ($i = 1 \ldots 5$) and one car fac-

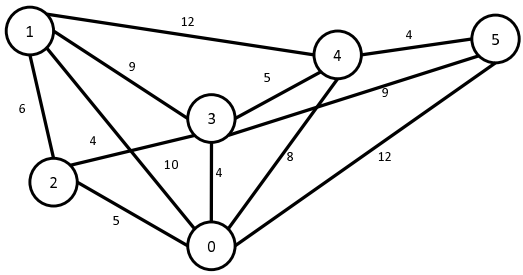


Fig. 2 Supplier network with transportation costs [4]

tory ($i = 0$). Suppliers 1, 2 and 3 have the same probability distribution for demand with the possible realization 1 and 2. $Pr(D_i = 1) = 0.6$ and $Pr(D_i = 2) = 0.4$. The suppliers 4 and 5 have the same possible realization 1 and 2, but a different probability distribution. $Pr(D_i = 1) = 0.4$ and $Pr(D_i = 2) = 0.6$. There are two vehicles with capacity $Q = 5$. The inventory cost h per unit is 0.2. The safety factor z is 1.0 and $\sigma_{hd}(a)$ is approximated by a as the value of the safety stocks linearly depends on a. The allowed probability that the capacity constraint is violated is $\alpha = 0.1$. The optimal solutions for the chance-constrained formulation are shown in Table 1. The solution without leveling is given for $a = 0$. This means the

Table 1 Chance-constrained solutions

a	**Route**	**Stops**	$\mathbf{C_{trans}}$	$\max \sum_i \Gamma_i$	$\mathbf{Pr}(\sum_i \Gamma_i > \mathbf{Q})$	$\mathbf{C_{ss}}$	$\mathbf{C_{total}}$
$a = 0$	1	1,2,3	24	6	0.064	0	24
	2	4,5	24	4	0	0	24
		Total	48			0	48
$0 \leq \frac{a}{2} < \frac{1}{3}$	1	1,2,3	24	$6 - \frac{3}{2}a$	0.064	$0.6 \cdot a$	$24 + 0.6 \cdot a$
	2	4,5	24	$4 - a$	0	$0.4 \cdot a$	$24 + 0.4 \cdot a$
		Total	48			a	$48 + a$
$\mathbf{\frac{a}{2} = \frac{1}{3}}$ (optimal solution)	1	1,2	21	3.33	0	0.267	21.267
	2	3,4,5	25	5	0	0.40	25.40
		Total	46			0.667	**46.667**
$\frac{1}{3} < \frac{a}{2} < \frac{3}{5}$	1	1,2	21	$4 - a$	0	$0.6 \cdot a$	$21 + 0.4 \cdot a$
	2	3,4,5	25	$6 - \frac{3}{2}a$	0	$0.4 \cdot a$	$25 + 0.6 \cdot a$
		Total	46			a	$46 + a$
$\frac{3}{5} \leq \frac{a}{2}$	1	1,2	21	2.8	0	$0.6 \cdot a$	$21 + 0.4 \cdot a$
	2	3,4,5	25	4.6	0	$0.4 \cdot a$	$25 + 0.6 \cdot a$
		Total	46			a	$46 + a$

order quantity Γ_i is identical to the demand D_i. Therefore there is no difference between order quantity and demand and thus $C_{ss} = 0$. The route with stops at 3, 4 and 5 is not chosen because the probability $Pr(\sum_i \Gamma_i > Q)$ for this route is 0.144 and

greater than the allowed probability $\alpha = 0.1$. If $\frac{a}{2}$ is increased to $\frac{1}{3}$ the solution does not change because the interval of the control band policy is too small to have a decisive effect on the maximum order quantity $\max \sum_i \Gamma_i$. However, at the same time more safety stock is needed to hedge against the increased demand variability. The optimal solution is $\frac{a}{2} = \frac{1}{3}$ with route 1 stopping at 1 and 2 and route 2 stopping at 3, 4 and 5 since the sum of the order quantity of route 2 is equal to the vehicle capacity Q and does not violate the stochastic capacity constraint. Furthermore the safety stock costs are less than the savings in the transportation costs C_{trans}. With increasing interval width the safety stock costs increase, but the transportation costs remain constant. This is due to the fact that the degree of leveling has no influence on vehicle routing. When $\frac{a}{2}$ reaches the value $\frac{3}{5}$, the maximum degree of leveling is achieved and the order quantity equals the expected demand $\mu_i \in \{1.4,\ 1.6\}$ for every order.

The optimal solution indicates that leveling the order quantities is advantegous compared to the solution with the maximum possible order quantities. This is because leveling saves more transportation costs than costs for safety stock are added.

4 Conclusion and Outlook

In this paper we presented a vehicle routing problem with the option to level order quantities. Therefore, we combined a leveling order policy with the SCVRP. The objective was calculating the optimal degree of leveling order quantities and the optimal routes to minimize transportation and safety stock costs. The new decision variable "interval width" was introduced into the chance-constrained formulation of the SCVRP. An example showed that a certain degree of leveling is advantageous with regards to the total costs.

For the automotive industry the decision about order quantities does not only depend on the demand. Likewise, other factors as lead time and actual inventory level are relevant. Furthermore a very simplified probability distribution of the demand was used in the example. A description of the probability distribution of the order quantities in dependence on the interval width is needed. The fact that the model is NP-complete and the large size of real problems requires a powerful heuristic.

References

1. Herbert Meyr. Supply chain planning in the german automotive industry. *OR Spectrum*, 26(4): 447–470, 2004.
2. Melda Ormeci, J. G. Dai, and John Vande Vate. Impulse control of brownian motion: The constrained average cost case. *Operations Research*, 56(3): 618–629, 2008.
3. Edward A. Silver, David F. Pyke, and Rein Peterson. *Inventory management and production planning and scheduling*. Wiley, New York, NY, 3 edition, 1998.
4. William R. Stewart and Bruce L. Golden. Stochastic vehicle routing: A comprehensive approach. *European Journal of Operational Research*, 14(4): 371–385, 1983.

Tactical Planning in Flexible Production Networks in the Automotive Industry

Kai Wittek, Thomas Volling, Thomas S. Spengler, and Friedrich-Wilhelm Gundlach

Abstract Today's automotive markets are characterized by high demand volatility. As a consequence, flexibility to re-allocate production volume within the production networks is required, to match the production capacity for specific models with market demand. The sustained re-allocation of models to production lines to adjust to market demand is no longer a vision, but virtually industry standard. Hence, model plant allocation decisions are no longer a one time long term decision but a more frequent mid term planning task, for which current strategic decision support models are inadequate. Modified planning models are required to handle the increased flexibility. A modeling framework will be presented.

1 Introduction and Motivation

Automotive original equipment manufacturers (OEMs) face a strong competition and rising differentiation of demand, requiring a high product variety. This variety increases market fragmentation, making demand less predictable. A highly dynamic market has to be matched by capital intensive production facilities with fixed capacities. To address this challenge, OEMs increase the flexibility of their production networks driven by technological advances. This encompasses a bottom up approach based on standardization as depicted in Figure 1.

Common part, module and platform strategies standardize to an increasing extent products from a production process point of view. This enables standardized tacts

Kai Wittek · Thomas Volling · Thomas S. Spengler
Institute of Automotive Management and Industrial Production, Technische Universität Braunschweig, Katharinenstr. 3, 38106 Braunschweig, e-mail: k.wittek | t.volling | t.spengler@tu-bs.de

Friedrich-Wilhelm Gundlach
Volkswagen AG, Brieffach 1653, 38436 Wolfsburg, e-mail: friedrich-wilhelm.gundlach@volkswagen.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_68,

Fig. 1 Achieving Flexibility in Automotive Production Networks

in final assembly across several models (e.g. Polo, Fabia) allowing for mixed model lines. Having standardized products and tacts, tools and equipment can be standardized admitting the definition of standardized plants with standardized processes. Combined, these plants constitute a production network. Due to standardization, models can now be re-allocated within the network. This results in networks with allocation flexibility, hence they are denoted flexible networks.

Flexibility is a hard to capture concept with various definitions and terms in the literature. A survey by [9] found more than 50 different terms. Here, flexibility refers to being able to produce multiple models at a single plant on a single final assembly line and being able to shift one model from one plant to another within the production network with limited or no capital investments.

The economic benefit of flexibility to shift production volume between plants is manifold. Lost sales are reduced, since capacity for a specific model can be adjusted by shifting peak demand to underutilized plants. Further, by aggregating demand on mixed model lines, the capacity utilization is leveled and increased. The option to re-allocate substitutes capacity expansions. However, when re-allocating, the product to plant allocation no longer is fixed in the mid to short term. This paper assumes the described scenario of flexible networks and the allocation problem moves from the long term to the mid and short term, posing new requirements for planning.

Literature on flexibility in the automotive industry is focused on plant flexibility [1] or chaining concepts [6] to generate flexibility. Product plant allocation in automotive networks is considered on a strategic level with deterministic models by [7, 3] or by [4, 2] with stochastic models. Current approaches assume a fixed product to plant allocation and do not consider re-allocation in the mid or short term. This does no longer match the realities encountered in the automotive industry, where flexibility to re-allocate is available. Further, the possibility to shift or defer demand between planning periods is not considered on a strategic level, where lead time aspects are not incorporated. These however, become relevant when re-allocating in the mid to short term.

The goal of this paper is to develop a modeling approach for the product plant allocation in flexible automotive production networks, taking re-allocation in the mid and short term as well as the option to shift and defer demand into account. The paper is structured as follows: First, the planning problem and problem situation will be described in detail in section 2. From this, modeling requirements will be derived. In section 3 a modeling approach fulfilling these requirements will be developed. The paper finishes with conclusions and an outlook in section 4.

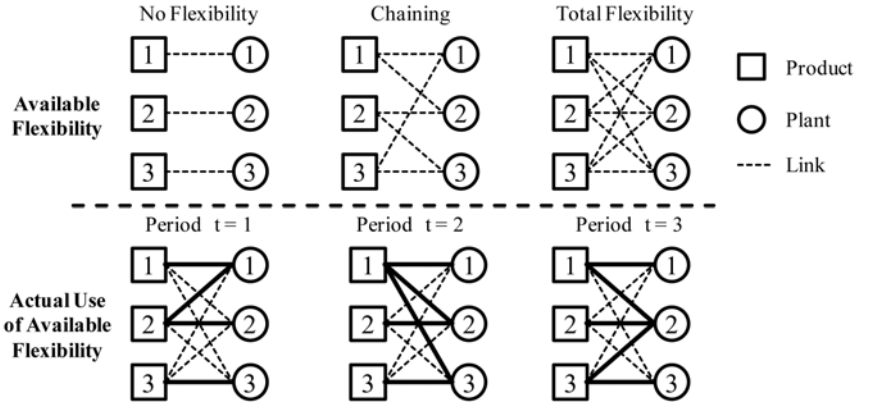


Fig. 2 Flexibility Configuration and Usage

2 Planning Problem

The mid term product plant allocation in flexible networks is dependent on the flexibility configuration defined by strategic network planning, providing the framework for any mid and short term shifts of product volume in the network. Different flexibility configurations are depicted in the upper part of Figure 2. Coming from dedicated facilities with no flexibility, the chaining concept [6] allowing for considerable flexibility with a limited number of links was developed. Nowadays, many automotive production facilities are already multi model capable and automotive production networks advance towards total flexibility. The product plant allocation has to use this given flexibility efficiently in the mid to short term. As depicted in the lower part of Figure 2, flexibility allows for re-allocation as time elapses. Product 2 is removed from plant 1 in period 2 and product 1 added to plants 2 and 3. When now a segment shift occurs, another re-allocations can take place.

With a mid to short term time horizon, further aspects become relevant. In addition to determining where to produce product volume, shifting or deferring demand volume has to be considered. It may be beneficial to shift production to an earlier period to reduce lost sales but at the price of incurring holding costs. Deferring to a later period may avoid expenditures for capacity expansions. This however, increases lead times and customers may be lost. In Figure 3 we exemplarily consider a decision situation of model plant allocation over four periods for two products to be allocated to two plants with fixed capacity. Demand for product 1 is deferred from period 1 to period 2, demand for product 2 in period 2 is shifted to period 1. Further, demand is re-allocated from plant 2 to plant 1 in periods 3 and 4. Whenever capacity is insufficient to fulfill demand, like in period 3, the difference in product value becomes relevant. It has to be decided which demand should be fulfilled from an economic perspective like in capacity control. This leads to a complex combination of planning problems to be addressed simultaneously.

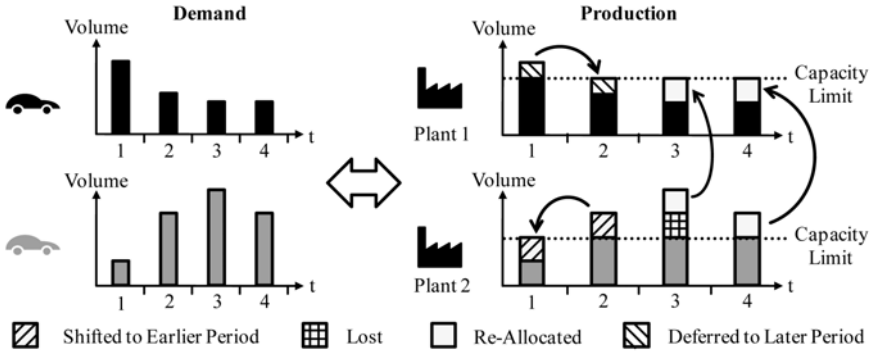


Fig. 3 Planning Problem

Since local to local production is on the rise and the majority of production volume stays within the economic region (e.g. European Union) where produced, they are the system boundary of the planning problem. This allows for omitting tariffs and logistics costs are less relevant. To summarize, a model for product plant allocation in flexible production networks has to fulfill the following requirements in extension of current strategic models: a) the time horizon is mid to short term b) products have different revenues c) re-allocation of products to plants is possible as time elapses d) production volume may be deferred or shifted in time resulting in holding costs or lost customers. An integrated model not only determining where to produce (allocation planning), but also when and what product to produce (master production planning), is required. A modeling approach will be presented below.

3 Modeling Approach

We consider an automotive OEM having multiple plants p and products (models) m. The focus is on the final assembly line and each product is assumed to have the same capacity consumption. Production volume $P_{m,p,t}$ has to be allocated in a multi period setting. The granularity of one period t is one month and the time horizon two years. Restrictions to the model plant allocation apply, defined by the matrix $R_{m,p}$ with entries $\in (0,1)$. Plant capacity can be increased by overtime incurring additional costs. Further, we assume that production for demand $D_{m,k,t}$ for product m in economic region k can be shifted to earlier periods, incurring holding costs. Production can in addition be deferred to later periods leading to lost customers. Reallocation of a model generates costs. A mixed integer linear programming (MILP) model with a time indexed formulation results. All variables are positively valued.

A shortfall based objective function as in [4] is inadequate for considering different product values and hence a monetary objective function as in [7, 3] was chosen. To account for temporal effects, it is discounted by the interest rate i. However,

since different product values i.e. revenues are of relevance, the net present value (NPV) over the cash in- and outflows is maximized and not only investments minimized.

The model is related to and partially based on the strategic model formulations of [7, 3, 2]. Hence, the focus in the model presentation is on extensions like shifting and deferring demand. The objective function (1) includes the following elements, where the term costs refers to actual costs i.e. actual cashflows:

$$\max \sum_t \frac{1}{(1+i)^t} [revenues_t \text{ - } production\ costs_t \text{ - } costs\ for\ extra\ shifts_t \text{ - } inventory\ holding\ costs_t \text{ - } re\text{-}allocation\ costs_t] \quad (1)$$

When deferring and shifting demand, the demand fulfillment has to be separated into shifted production $SP_{m,p,t^{in},t}$, deferred production $DP_{m,p,t^{in},t}$ as well as production $P^d_{m,p,t}$ produced in t for demand in t. Shifted and deferred production are produced in t^{in} for demand fullfillment in t. Limits on shifting and deferring are given implicitly by incurring holding costs for shifted demand and by loosing customers when deferring. The rate of loosing customers is $x_{m,k}$ per period. Discounting the cashflows in the objective function further provides implicit limits on the economic attactiveness of deferring and shifting. The general expression is given in (2):

$$demand \geq production\ in\ t\ for\ t + deferred\ production + shifted\ production \quad (2)$$

In (3) the detailed mathematical expression is given for all k,m,t:

$$D_{m,k,t} \geq \sum_p P^d_{m,p,t} + \sum_p \sum_{t^{in}=1}^{t-1} SP_{m,p,t^{in},t} + \sum_p \sum_{t^{in}=t+1}^{T} DP_{m,p,t^{in},t} * \left(1+x_{m,k}\right)^{t^{in}-t} \quad (3)$$

Of particular interest is deferred production, since whenever deferring production, customers are lost, the more, the further production is deferred. This is modelled by an inverse present value, where $t^{in}-t$ describes the time span, demand is deferred. An example would be: $t^{in}-t=2$, $x_{m,k}=0.1$, and after production in t for t and shifted production, demand for 121 customers is still unfulfilled. Deferring it by 2 periods results in 21 lost customers and a deferred production of 100.

The total production volume $P_{m,p,t^{in}}$ of model m in plant p in period t^{in} results in:

$$P_{m,p,t^{in}} = P^d_{m,p,t^{in}} + \sum_{t=t^{in}+1}^{T} SP_{m,p,t^{in},t} + \sum_{t=1}^{t^{in}-1} DP_{m,p,t^{in},t} \quad (4)$$

To allow for restrictions on re-allocations as well as for incurring re-allocation costs, re-allocations need to be detected. [5] applies a re-allocation variable, however only for determining adding products to a line incurring investments. In flexible networks however, the focus has to be on operational costs, which applies for adding and removing products. This is enabled by $RA_{m,p,t} \in (0,1)$ stating if a re-allocation (added or removed) of product m is conducted in plant p in period t. It is based

on the allocation variable $AV_{m,p,t}$, which states if product m is allocated to plant p in period t. Taking the sum of the allocation variable from period t and the following $t+1$ and using two auxiliary variables $RA^1_{m,p,t}$; $RA^2_{m,p,t}$ both $\in (0,1)$ the value of $RA_{m,p,t}$ can be determined. E.g. adding a product m to plant p in $t+1$: $AV_{m,p,t} = 0$ and $AV_{m,p,t+1} = 1$ results in $RA^1_{m,p,t}$; $RA^2_{m,p,t}$ taking the value zero, forcing $RA_{m,p,t}$ to take the value one. The other three cases are likewise.

$$RA_{m,p,t} = 1 - \left(RA^1_{m,p,t} + RA^2_{m,p,t}\right) \qquad \forall m,p,t = 1,...,T-1 \qquad (5)$$

$$AV_{m,p,t} + AV_{m,p,t+1} + RA^1_{m,p,t} - RA^2_{m,p,t} = 1 \qquad \forall m,p,t = 1,...,T-1 \qquad (6)$$

Several other restrictions as well as initialization and stopping criteria apply, which could not be given in the scope of this paper. Since early allocation decisions have a limited impact on later decisions due to re-allocations, the above model is adequate to address the planning situation of the outlined flexibility scenario.

4 Conclusion and Outlook

In this article the product plant allocation in flexible networks requiring new approaches is addressed. An approach integrating operative, tactical and strategic requirements is presented. Future work includes model extensions e.g. based on lot sizing as well as comprehensive numerical analysis using real world data. Further, the planning problem is to be embedded in the hierarchical supply chain planning [8] of the automotive industry, including a rolling horizon.

References

1. G. Askar and J. Zimmermann. Optimal usage of flexibility instruments in automotive plants. In K.-H. Waldmann and U.M. Stocker, editors, *Operations Research Proceedings 2006*, pages 479–484, Berlin, Heidelberg, 2007. Springer Verlag.
2. R. Bihlmaier, A. Koberstein, and R. Obst. Modeling and optimizing of strategic and tactical production planning in the automotive industry under uncertainty. *OR Spectrum*, 31(2): 311–336, 2009.
3. B. Fleischmann, S. Ferber, and P. Henrich. Strategic planning of BMW's global production network. *Interfaces*, 36(3): 194–208, 2006.
4. D. Francas, M. Kremer, S. Minner, and M. Friese. Strategic process flexibility under lifecycle demand. *International Journal of Production Economics*, 121(2): 427–440, 2009.
5. P.R. Gneiting. *Supply Chain Design fuer modulare Fahrzeugarchitekturen*. Eidgenoessisch Technische Hochschule, Zuerich, 2009.
6. W. C. Jordan and S. C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4): 577–594, 1995.
7. S. Kauder and H. Meyr. Strategic network planning for an international automotive manufacturer: Balancing flexibility and economical efficiency. *OR Spectrum*, 31(3): 507–532, 2009.
8. H. Meyr. Supply chain planning in the german automotive industry. *OR Spectrum*, 26(4): 447–470, 2004.
9. A. K. Sethi and S. P. Sethi. Flexibility in manufacturing: a survey. *The International Journal of Flexible Manufacturing Systems*, 2(4): 289–328, 1990.

Coordination by Contracts in Decentralized Product Design Processes – Towards Efficient Compliance with Recycling Rates in the Automotive Industry

Kerstin Schmidt, Thomas Volling, and Thomas S. Spengler

Abstract Design processes in the automotive industry are distributed over various companies. If fixed-price contracts are used to coordinate such collaborations, inefficiencies in the design process occur due to existing uncertainties and differing objectives of the partners. To make decentralized design processes more flexible and to reduce inefficiencies, we introduce the conceptual design of incentive contracts and apply it to the case of the compliance with recycling rates in the automotive industry.

1 Motivation

During the design phase all future stages of the product's lifecycle, production, usage as well as recycling and disposal, are influenced to a large degree. For this reason, the legislator tries to impair the product design. In the automotive industry legal requirements, like the EU Directive 2000/53/EG and its implementation into national law (AltfahrzeugG) as well as the EU Directive 2005/64/EG, pose an important framework for sustainable design processes. For example, since December 2008 automotive manufacturers (OEMs) are obligated to demonstrate the reuse and recovery rate of at least 85 % and the reuse and recycling rate of at least 80 % when the type-approval of a new vehicle is done. From January 2015 these rates shall be increased to a minimum of 95 % and 85 %, respectively.

Currently, in the automotive industry, design processes are distributed over companies. Hence, the components of a vehicle are not designed centrally at the OEM, but at a variety of specialized suppliers. In this distributed design process the OEM acts as an integrator and bears responsibility for the compliance with statutory recycling rates. However, the recycling rates depend on the characteristics of each

Kerstin Schmidt · Thomas Volling · Thomas S. Spengler
Institute of Automotive Management and Industrial Production, Technische Universität Braunschweig, Katharinenstr. 3, 38106 Braunschweig, e-mail: kerstin.schmidt | t.volling | t.spengler@tu-bs.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_69,

component of the vehicle and thus on the design effort of the suppliers. Since the partners of this decentralized design process are usually legally and economically independent companies, the cooperation is regularized by contractual agreements. If inappropriate contract structures are used, existing uncertainties and differing objectives of the partners can lead to inefficiencies in the design process. If statutory recycling rates are exceeded, additional development costs result for the suppliers, without adding any additional utility for the OEM. If recycling rates are violated, the new vehicle cannot be certified to the intended date and additional design effort of suppliers is required. Thus, the economic risk for suppliers and OEM increases. Overcoming these difficulties requires improved coordination between the partners before and during the design process.

The economics of contracts have been studied in a variety of quantitative academic contributions [2, 7]. Particularly in supply chain management settings of independent OEMs and suppliers are analyzed. A comprehensive review of contract analysis in this field is provided by [7]. In the context of supply chain management, a central phenomenon causing inefficiencies is uncertainty. In these situations the flexibility to change a chosen course of action is usually associated with better performance [1]. Accordingly, uncertainty and thus risk is being shared with respect to the contractual agreement. Of particular interest for the research presented are flexible contract types [1, 6]. However, differences between supply chain management and design processes currently prevent an easy adoption of the mentioned flexible contracts. First approaches to the flexibilization of contracts in design processes are given by [3, 4, 5]. Since these contributions focus on a general analysis of the effectiveness of incentive contracts, a formal modeling approach for the analysis of contracts and their coordination ability in distributed design processes is still missing.

To this end, the conceptual design of incentive contracts in decentralized product design processes is presented and applied to the case of the compliance with recycling rates in the automotive industry. The rest of the paper is structured as follows. After modeling and analyzing decentralized design processes in the next section, we present an approach to improve decentralized design processes in section 3, before we finally draw our conclusions and give a short outlook on our future research.

2 Modeling and Analysis of Decentralized Design Processes

In this section we will study current decentralized design processes in the automotive industry. In the first step we will present, how design processes can be modeled. In the second step we will concentrate on the analysis of these processes.

2.1 Model Description

We consider three independent actors with full information, one OEM and two suppliers. The two suppliers each develop one component of a new vehicle for the OEM, for example the instrument panel and the underbody protection. To achieve different recycling-relevant specifications $s_i \in [s_{i,min}, s_{i,max}]$ of each component, with $i = 1,2$ indicating the two different suppliers, there exist various technical possibilities, like reducing or substituting nonrecyclable materials or using detachable connections. Each of these possibilities goes along with a specific design effort and thus specific development costs for the suppliers. Different combinations of the recycling-relevant specifications of both components can lead to the compliance with the statutory recycling rates for the new vehicle. We assume that these combinations, that lead to the compliance with recycling rates, can be described approximatively by the functional relationship $s_2 = r(s_1)$.

Since uncertainties exist in the design process, the component specifications s_i of the suppliers, resulting at the end of their design process, cannot be predicted with certainty, when the contract is concluded. It is assumed, that the probability of achieving a certain value of the component's specification is uniformly distributed in the interval $[s_{i,min}, s_{i,max}]$. However, the suppliers can still influence the development results s_i with their efforts w_i, e.g., number of employees working on the project. Due to practical relevance it has to be assured that the expected component specification is zero, when the effort is zero. Furthermore, the interval nearly shall remain the same size and the component's specification shall converge to a technological threshold with increasing effort. According to these requirements, the coherence between component specification and effort is implemented as follows: The interval limits depend on w_i: $s_{i,min} = ln(a_i \cdot w_i + 1)$ and $s_{i,max} = ln(b_i \cdot w_i + 1)$, with $b_i > a_i > 0$. Thus, a greater effort shifts the interval limits to the right, so that an increase in effort leads to a higher expected value of s_i.

The expected utility of the OEM depends on the compliance with the recycling rates of the new vehicle and thus on the expected specifications s_i of the suppliers' components. If the recycling rates are met, the utility of the OEM is one, otherwise it is reduced by his opportunity costs. To model the binary-character of the utility function in a continuously and differentiable manner, it is approximated by a Sigmoid-function, which is normalized to $[0,1]$. Through its turning point the function $r(s_1)$ is running, so that the substitutional interdependencies between the two component specifications are taken into account. Since the efforts w_i of the suppliers are stochastically independent, the existing uncertainties get included by multiplying the Sigmoid-function with the combined density function $P(s_1, s_2)$ and integrating on the two intervals, with the interval limits depending on the efforts of the suppliers. This results in the following expected utility function n of the OEM, with d specifying the slope of the turning point, $d \to \infty$:

$$n(w_1, w_2) = \int_{s_{1,min}}^{s_{1,max}} \int_{s_{2,min}}^{s_{2,max}} \frac{1}{1 + \exp(d \cdot r(s_1)) \cdot \exp(-d \cdot s_2)} \cdot P(s_1, s_2) d(s_2) d(s_1) \qquad (1)$$

The development costs of the suppliers depend on their efforts w_i and thus can be described by the function $c_i(w_i)$, normalized to $[0,1]$.

2.2 Illustrative Model Analysis

Centralized Setting. In the centralized setting (c) one decision maker, who has full information, develops and integrates both components. Hence, his aim is to maximize the total profit Π of the design process. His profit function results from the expected utility n minus the development costs c_i as follows:

$$\Pi(w_1, w_2) = n(w_1, w_2) - c_1(w_1) - c_2(w_2) \tag{2}$$

The optimal efforts $\bar{w}_{i,c}$ as well as the resulting expected component specifications $\bar{s}_{i,c}$ result from the analytical determination of the extreme point of the profit function Π. These results provide as a benchmark for the analysis of the coordination ability of different kinds of contracts in the decentralized setting.

Decentralized Setting. In the decentralized setting (fp), the decision situation of the suppliers is considered, when collaboration is based on a fixed-price contract. Currently, fixed-price contracts are most commonly used in the automotive industry. In these contracts the OEM determines certain recycling-relevant specifications $\hat{s}_i$ for the components that have to be fulfilled by the suppliers. Based on these parameters, the OEM then specifies the transfer payment determining the amount to be paid by him to the suppliers. In this fixed-price contract, a component specification equal or higher than the required specification leads to a payment corresponding to the development costs of the suppliers, while a specification lower than the required specification leads to a payment corresponding to the development costs reduced by the penalty costs. Since the realized specifications, and thus the resulting transfer payment, depend on the suppliers' effort, the expected transfer payment functions arise as follows: To model the binary-character of the transfer payment in a continuously and differentiable manner, it is approximated by a Sigmoid-function, which is normalized to $[0,1]$. The turning point of the Sigmoid-function is described by the required component specification $\hat{s}_i$. The existing uncertainties are included by multiplying the Sigmoid-function with the density function $P(s_i)$ and integrating on the interval, with the interval limits depending on the effort of the supplier. The expected transfer payment function $t_{i,fp}^{\hat{s}_i}$, with m specifying the slope in the turning point, $m \to \infty$, results as follows:

$$t_{i,fp}^{\hat{s}_i}(w_i) = \int_{s_{i,min}}^{s_{i,max}} \frac{c_i(\hat{s}_i)}{1 + \exp(m \cdot \hat{s}_i) \cdot \exp(-m \cdot s_i)} \cdot P(s_i) d(s_i) \tag{3}$$

We assume that the OEM chooses the required component specification $\hat{s}_i$ according to the optimal specification $\bar{s}_{i,c}$ in the central case. The profit functions $\pi_{i,fp}^{\hat{s}_i}$ of the suppliers are based on the transfer payment functions minus their development costs. The suppliers then determine their effort w_i by maximizing their profit

function $\pi_{i,fp}^{\hat{s}_i}$:

$$\pi_{i,fp}^{\hat{s}_i}(w_i) = t_{i,fp}^{\hat{s}_i}(w_i) - c_i(w_i) \tag{4}$$

The analysis shows, that the optimal efforts $\bar{w}_{i,fp}^{\hat{s}_i}$ and thus the expected component specifications $\bar{s}_{i,fp}^{\hat{s}_i}$ are higher than in the centralized setting. Due to a fixed-price contract the suppliers choose a higher effort than necessary to achieve at least the design specifications $\hat{s}_i$ and thus get the agreed transfer payment with certainty. Consequently, inefficiencies in the design process occur and increased costs and overdesigned components are resulting. Hence, new types of contracts are necessary to improve the cooperation between OEM and suppliers in distributed design processes.

3 Improving Decentralized Design Processes

In this section we will present the conceptual design of an incentive contract to make decentralized design processes in the automotive industry more flexible and thus less inefficient. In this decentralized setting (in), the decision situation of the suppliers is considered, when collaboration is based on an incentive contract. In incentive contracts not a certain design specification is set by the OEM, but an interval on design specifications. For every specification within the interval the suppliers get an agreed transfer payment $h_i(s_i)$ (e.g., the transfer payment increases linear with respect to the achieved specification). Additionally, the suppliers are able to select a component specification within this interval and thus an effort to achieve this specification, which fits best to their structure of development costs and hence to their uncertainties. Since the realized specifications, and thus the resulting transfer payments, depend on the suppliers' effort, the expected transfer payment function arises as follows: The existing uncertainties are included by multiplying the transfer payment $h_i(s_i)$ with the density function $P(s_i)$ and integrating on the interval, with the interval limits depending on the effort of the supplier. The expected transfer payment function $t_{i,in}^{h_i}$ results as follows:

$$t_{i,in}^{h_i}(w_i) = \int_{s_{i,min}}^{s_{i,max}} h_i(s_i) \cdot P(s_i) d(s_i) \tag{5}$$

The profit functions $\pi_{i,in}^{h_i}$ of the suppliers are based on the transfer payment functions minus their development costs. The suppliers then determine their effort w_i by maximizing their profit function $\pi_{i,in}^{h_i}$:

$$\pi_{i,in}^{h_i}(w_i) = t_{i,in}^{h_i}(w_i) - c_i(w_i) \tag{6}$$

The analysis shows, that the optimal efforts $\bar{w}_{i,in}^{h_i}$ and the resulting expected component specifications $\bar{s}_{i,in}^{h_i}$ are higher than in the centralized setting, but lower than in

the decentralized setting with fixed-price contract. Thus, using incentive contracts enables to reduce the overdesign of the components and to make a contribution to the coordination of decentralized design processes. The reason for the reduced overdesign of the suppliers is their higher flexibility and the missing risk to get no transfer payment. Due to the higher flexibility for the suppliers the risk of the OEM increases, since the chosen efforts of the suppliers can lead to the non-compliance with the statutory recycling rates. Therefore it is necessary for him to set the contract parameters h_i advisedly.

4 Conclusion and Outlook

In this paper we presented a formal modeling approach for decentralized product design processes in the automotive industry. In the first step we analyzed a centralized setting and compared the results to a decentralized setting with a fixed-price contract. The analysis showed that fixed-price contracts lead to inefficiencies. Hence, in the second step we introduced the conceptual design of incentive contracts to make decentralized design processes in the automotive industry more flexible. Thus, inefficiencies could be decreased by dividing the development risk between the OEM and his suppliers. Our future research will concentrate on the integration of the decision situation of the OEM in the decentralized setting.

References

1. R. Anupindi and Y. Bassok. Supply contracts with quantity commitments and stochastic demand. In *Quantitative models for supply chain management*, pages 197–232. Kluwer Academic Publishers, 1999.
2. P. Bolton and M. Dewatripont. *Contract theory*. The MIT Press, Cambridge, 2005.
3. J. Kruse, C. Thomsen, R. Ernst, T. Volling, and T.S. Spengler. Introducing flexible quantity contracts into distributed SoC and embedded system design processes. In *Proceedings of the conference on Design, Automation and Test in Europe*, pages 938–943, 2005.
4. J. Kruse, C. Thomsen, R. Ernst, T. Volling, and T.S. Spengler. Towards flexible systems engineering by using flexible quantity contracts. In *Proceedings of the conference on Automation, Assistance and Embedded Real Time Platforms for Transportation*, pages 480–488, 2005.
5. J. Rox, K. Schmidt, A. Winter, T.S. Spengler, and R. Ernst. Estimating and mitigating design risk in a flexible distributed design process. *IEEE Embedded Systems Letters*, 2(2): 35–38, 2010.
6. A.A. Tsay. The quantity flexibility contract and supplier-customer incentives. *Management science*, 45(10): 1339–1358, 1999.
7. A.A. Tsay, S. Nahmias, and N. Agrawal. Modeling supply chain contracts: A review. In *Quantitative models for supply chain management*, pages 299–336. Kluwer Academic Publishers, 1999.

Evaluating Procurement Strategies under Uncertain Demand and Risk of Component Unavailability

Anssi Käki and Ahti Salo

Abstract In high-technology manufacturing, failure to access non-commodity components can interrupt production, because there are typically no substitute suppliers. Moreover, demand and supply uncertainties can be interdependent, because supplier problems are more probable when demand is high. Most earlier risk management models treat these uncertainties as independent risk factors. We present a framework where judgmental information is captured with a scenario tree and a stochastic decision model is used to evaluate alterative procurement strategies consisting fixed quantity contracts for cost minimization and flexible quantity contracts for risk management.

1 Introduction

Demand and supply uncertainties are of growing concern in supply chain management [3, 9]. Many examples, such as Nokia's success after semiconductor plant fire [8] or Cisco Systems' excess inventory write-offs during an economic downturn [7] have highlighted risks caused by these uncertainties. However, as [9] states: *the challenges are great, but so are the opportunities*. For instance, Hewlett-Packard has realized more than $425 million in cost savings due to the implementation of effective procurement strategies [11].

In supply risk management, the most prevalent strategy is supply diversification, e.g. [1], or backup from SPOT-market [10]. However, there is often only one supplier for a sophisticated high-technology component; e.g. [4] develop a single-supplier model with multiple components and products, with independent supply capability distributions, and stationary random demand. Overall, recent research indicates that quantification and management of risks can provide profitable, even if

Anssi Käki and Ahti Salo
Aalto University, School of Science and Technology, Systems Analysis Laboratory, P.O. Box 11100, FIN-00076, e-mail: anssi.kaki@tkk.fi

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_70,

existing risk management models have typically considered only stationary and independent uncertainties.

In contrast to these models, we consider dependent and non-stationary demand and supply uncertainties. Specifically, we consider a consumer electronics manufacturer, whose main sources of risks are: i) volatile end product demand, ii) supply of customized components with long lead times and iii) *strong interdependencies* between demand and supply (i.e. when demand grows, supply problems are more likely). In this setting, we capture these uncertainties with scenario trees and develop a stochastic programming model for the evaluation of alternative procurement strategies.

2 The Model

Consider a manufacturer who faces a three-stage cost minimization problem. The procurement alternatives are i) fixed component-specific orders at a low unit cost and ii) capacity reservations, which offer flexibility (see [10] for a portfolio approach with similar contracts). Through capacity reservation, the manufacturer can delay the decision on *what to order*, because it can decide later what components are needed in production, and *how much to order* subject to the limits of capacity reservation. We assume that the unit cost for unused capacity is smaller than the fixed order cost per unit, but the total cost of option (i.e. reservation + execution cost) exceeds the fixed order cost. Inventory holding costs and shortage costs are also considered.

Our approach is adapted from [5], who present a decision making process with three main phases: i) gathering and processing of data and judgmental knowledge, ii) creating a coherent set of scenarios, and iii) building and solving the stochastic decision model.

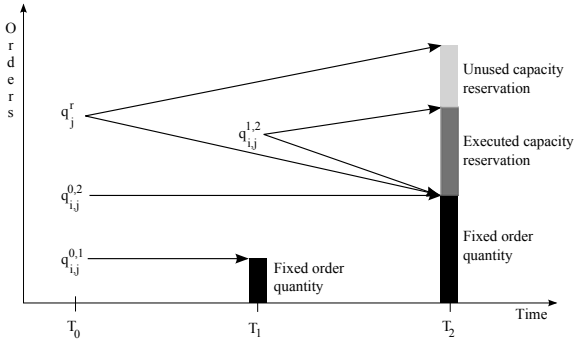


Fig. 1 The decision process has two stages: **1)** how much to reserve capacity q^r_j from supplier j and order fixed amounts $q^{0,1\ldots2}_{i,j}$ of component i from supplier j at initial stage and **2)** after initial demand realization, how much to execute capacity $q^{1,2}_{i,j}$ for each component i from supplier j to meet the final demand.

Our model has three time stages that cover the component life cycle: initial stage T_0 , when fixed quantity orders and capacity reservations are placed, early stage T_1 ,

when the early demand is observed and the decision to utilize capacity is made and final stage T_2, when the final demand is observed and salvage and stock-out costs are realized. The schematic description of model is in Figure 1.

Table 1 Notation

Decision variables	Unit costs	Cost functions	Other
$q_{i,j}^{0,t}$: Fixed order qty	$c_{i,j}^{o}$: Fixed order	f_c^0 : Initial stage	D_i^t : Demand
q_j^r : Reserved capacity qty	$c_{i,j}^{r}$: Capacity reservation	f_c^1 : Early stage	λ_j^t : Yield factor (0%...100%)
$q_{i,j}^{1,2}$: Executed capacity qty	$c_{i,j}^{e}$: Capacity execution	f_c^2 : Final stage	I_i^t : Inventory
	c_i^h : Holding		N_c : Amount of components
	c_i^s : Shortage		N_s : Amount of suppliers

In this notation, i stands for component, j for supplier and t for time (stage).
For all x_j^i, we note all variables over j with $\mathbf{x^i}$ and variables over i and j with $\mathbf{x}$.

Using the notation in Table 1, the stochastic procurement problem is:

$$\min_{\mathbf{q^r},\mathbf{q^0}} \left\{ f_c^0 + \mathrm{E}_1[\min_{\mathbf{q^1}} \left\{ f_c^1 + \mathrm{E}_2[f_c^2] \right\}] \right\} \tag{1}$$

$s.t.$

$$I_i^1 = [\sum_{j=1}^{N_s} \lambda_j^1 q_j^{0,1} - D_i^1]^+, i = 1 \ldots N_c \tag{2}$$

$$I_i^2 = [I_i^1 + \sum_{j=1}^{N_s} (\lambda_j^1 q_j^{0,2} + \lambda_j^2 q_j^{1,2}) - D_i^2]^+, i = 1 \ldots N_c \tag{3}$$

$$\sum_{i=1}^{N_c} q_{i,j}^{1,2} \leq q_j^r, j = 1 \ldots N_s \tag{4}$$

The cost functions $f_c^0(\mathbf{q^r}, \mathbf{q^0}, \mathbf{c^o}, \mathbf{c^r})$, $f_c^1(\mathbf{q^1}, \mathbf{I^1}, \mathbf{c^e}, \mathbf{c^h}, \mathbf{c^s})$ and $f_c^2(\mathbf{I^2}, \mathbf{c^h}, \mathbf{c^s})$ are linear expressions of costs and quantities. At the initial stage, the costs comprise fixed orders and reservations, at the early stage of option executions and inventory costs and at the final stage of inventory costs only. In the objective function (1), the expected cost E_2 is conditional and depends on the outcome of early stage random variables. The equality constraints (2) and (3) capture the component inventories at the early and final stages. Constraint (4) ensures that the quantity of capacity based orders is less than capacity reserved. All quantities q are non-negative integers.

The development of the *scenario tree* covers two main phases: i) *data preprocessing*, which typically includes analysis of historical data and processing of expert information and ii) *scenario creation*, where the scenario tree with desired statistical properties is constructed. Data preprocessing synthesizes relevant information to a joint distribution with statistical properties that match judgmental statements.

For example, we could have cumulative distributions both for demand and supply uncertainties for all components and suppliers, along with their correlations. We generate scenarios with *the moment-matching method* [6] by minimizing the distance between statistical moments of the underlying distribution and those of the scenario tree.

3 Numerical Example

We illustrate the model with a numerical example consisting of two suppliers and two components. First, we constructed a cumulative distribution for the demand of both components by approximating the percentiles of sales volumes that correspond to a typical demand pattern of a consumer electronics product (e.g., for 0% we set 10 000 units, for 50% 20 000, for 95% 100 000 and for 100% 1 000 000). From the empirical distribution, we estimated the first three moments for demand. The two suppliers had different characteristics, one "reliable" (expected reliability E[λ] = 97%) and one "cheap" (E[λ] = 96% and 10% less costs). The suppliers were characterized by the first two moments of reliability. The interdependency between uncertainties was introduced by specifying two scenarios with either strong interdependencies (correlation of 0.8 or higher) or weak interdependencies (correlation of less than 0.5). Finally, the following costs for both components were set: ordering cost (1.00, or 0.90 from "cheap" supplier), reservation cost (0.80-0.72), execution cost (0.70-0.63), holding cost (0.30) and shortage cost (5.00).

Based on the assumptions, two scenario environments were created: 1) *coupled* with strong interdependencies, indicating a strong positive correlation between component demands and a strong negative correlation between component demand and supplier capability and 2) *uncoupled* with weak interdependencies, meaning relatively small correlations between demands and capabilities (both within and between). The trees were constructed iteratively using visual inspection, which limits the size of trees. That is, our "moment matching" was not based on optimization, but on a heuristic where "close enough" scenario trees were acceptable. An example of two scenario trees is presented in Figure 2.

The stochastic procurement problem was solved for these scenario trees. The optimization results (see Table 1) indicate that strong dependencies lead to approximately 4% higher expected costs to the manufacturer. Also, dependencies seem to cause greater risks, because variation and range of costs are both smaller in the case of *uncoupled* scenario. An analysis of the cost structure showed that both order (12.9%) and reservation (6.5%) costs were higher for the *coupled* scenario, and the maximum shortage cost was over three times more than in the *uncoupled* scenario. This indicates that the worst-case risks due to interdependencies do matter. These results should be interpreted with caution, because the number of scenario paths was relatively small.

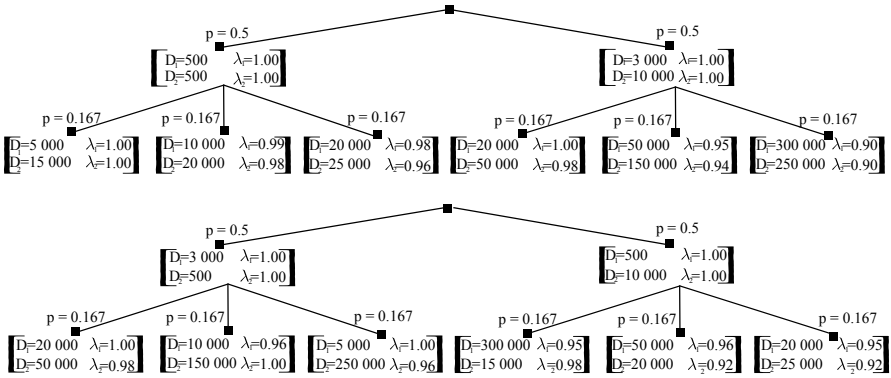


Fig. 2 Scenario trees for testing the model. **Top**: *coupled* with strong interdependencies between demand and supplier capability. **Bottom**: *uncoupled* with relatively small interdependencies both within and between demand and supply.

Table 2 Results of optimization

Scenario	E[Cost]	Std[Cost]	Min[Cost]	Max[Cost]
Coupled	506 623	747 003	174 288	2 030 350
Uncoupled	485 970	513 738	162 523	1 427 561

4 Discussion

We have studied the impact of interdependent uncertainties in high-technology procurement. Current procurement risk management models often assume stationary and independent uncertainties. Based on our initial discussions with a case company, these assumptions are unrealistic especially in consumer electronics manufacturing. The results of our optimization model suggest that interdependencies can be significant; if demands of products are positively correlated and/or the demand and supply capabilities are negatively correlated, worst-case procurement risks can be particularly significant.

In the future, we will explore systematic scenario generation with different dependency structures. More in-depth analysis is required to assess which interdependencies are particularly critical and how changes in the cost structure of suppliers, for instance, would affect overall procurement costs. We believe that our three-period model can be readily extended to more periods and also other to time spans than presented. Other promising extensions include more accurate models of supply uncertainty and the recognition of disruptive risks.

Recent surveys [13, 12] suggest that there is a gap between academics and supply chain practitioners in the use of quantitative models that consider uncertainties. There are some promising examples of using such models in procurement risk management, for instance [2, 11], but risk quantification is still not common in practice.

We see that adopting state-of-the-art risk management models that account for uncertainties and their dependencies is a key step in advancing procurement practices. We plan to study this proposition by applying our framework in a case study at a mobile phone manufacturer.

References

1. Awi Federgruen and Nan Yang. Selecting a portfolio of suppliers under demand and supply risks. *Operations Research*, 56(4): 916–936, 2008.
2. Marshall Fisher and Ananth Raman. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, 44(1): 87–99, 1996.
3. Marshall L. Fisher, Janice H. Hammond, Walter R. Obermeyer, and Ananth Raman. Making supply meet demand in an uncertain world. *Harvard Business Review*, 72(3): 83–93, 1994.
4. Haresh Gurnani, Ram Akella, and John Lehoczky. Supply management in assembly systems with random yield and random demand. *IIE Transactions*, 32(8): 701–714, 2000.
5. Ronald Hochreiter and Georg Ch. Pflug. Financial scenario generation for stochastic multistage decision processes as facility location problems. *Annals of Operations Research*, 152(1): 257–272, 2007.
6. Kjetil Hoyland and Stein W. Wallace. Generating scenario trees for multistage decision problems. *Management Science*, 47(2): 295–307, 2001.
7. Bill Lakenan, Darren Boyd, and Ed Frey. Why Cisco Fell: Outsourcing and Its Perils. *Strategy & Business*, 24: 1–12, 2001.
8. Almar Latour. Trial by fire: A blaze in Albuquerque sets off major crisis for cell-phone giants. In Wall Street Journal (January 29), 2001.
9. Hau L. Lee. Aligning supply chain strategies with products uncertainties. *California Management Review*, 44(3): 105–119, 2002.
10. Victor Martinez-de-Albeniz and David Simchi-Levi. A portfolio approach to procurement contracts. MIT Sloan School of Management Paper 188, Available at http://ebusiness.mit.edu/research/papers/188_DSlevi_PortfolioApproach.pdf, 2003.
11. Venu Nagali, Jerry Hwang, David Sanghera, Matt Gaskins, Mark Pridgen, Tim Thurston, Patty Mackenroth, Dwight Branvold, Patrick Scholler, and Greg Shoemaker. Procurement risk management (PRM) at Hewlett-Packard company. *Interfaces*, 38(1): 51–60, 2008.
12. David Peidro, Josefa Mula, Raúl Poler, and Francisco-Cruz Lario. Quantitative models for supply chain planning under uncertainty: a review. *The International Journal of Advanced Manufacturing Technology*, 43(3–4): 400–420, 2009.
13. ManMohan S. Sodhi and Christopher S. Tang. Modeling supply-chain planning under demand uncertainty using stochastic programming: A survey motivated by asset liability management. *International Journal of Production Economics*, 121(2): 728–738, 2009.

The Effects of Wholesale Price Contracts for Supply Chain Coordination under Stochastic Yield

Karl Inderfurth and Josephine Clemens

Abstract Coordinating the decisions of individual businesses in a supply chain (SC) can reduce efficiency losses which occur when companies act solely towards individual optimization. Incentives which ensure that all parties benefit from coordination can be set through contracts containing parameters of transfer payments. From research on SC coordination in many decision fields we know that a simple wholesale price (WP) contract will not coordinate due to the so-called double marginalization effect (see [1]). However, considering a SC exposed to yield randomness at the supplier side, it can be shown that also for the WP contract coordination property might be given depending on the specific SC environment. This aspect will be analyzed in detail.

1 Introduction

Uncertainty in SCs can occur in various forms, e.g. in form of demand and supply uncertainties. Random supply can exist due to different reasons such as production process risks or imperfect input material. Consequently, production yield is uncertain. While random demand can be found in almost all industries, supply uncertainty occurs more frequently in agricultural, chemical or semiconductor production than in other industries. The context of ordering under random supply has been studied extensively (see [8]) while stochastic production yield in the SC coordination context has not received much attention so far. In this case, appropriate contracts have to account for this kind of randomness. Several contract types have been recommended in literature which ensure coordination (see [2], [3]). The simple WP contract, however, is usually known to not coordinate due to double marginalization. This article, though, shows that the specific production setting determines whether the simple WP contract can coordinate the SC or not.

Karl Inderfurth, e-mail: karl.inderfurth@ovgu.de
Josephine Clemens, e-mail: josephine.clemens@ovgu.de
Faculty of Economics and Management, Otto-von-Guericke University Magdeburg

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_71,

2 The Supply Chain Model

Considering a serial SC with one buyer (indicated by subscript B) and one supplier (indicated by subscript S), the supplier produces at unit cost c and sells to the buyer at unit wholesale price w. The product is finally sold by the buyer at unit retail price p for fulfilling a deterministic demand D. Due to the fact that the supplier's production process is exposed to random yield the buyer faces a random fulfillment level of his order. It is assumed that the production yield is stochastically proportional, i.e. the usable production output $\tilde{Y}$ is a fraction of the production input Q such that $\tilde{Y} = \tilde{Z} \cdot Q$, where $\tilde{Z}$ is the random yield rate with *pdf* $f(\cdot)$, *cdf* $F(\cdot)$ and mean μ_Z.

2.1 SC Coordination under Random Yield without External Procurement Option

In this scenario, the buyer orders the quantity q while facing deterministic demand D. The supplier has only a single production run which is exposed to randomness. Thus, his decision is solely on the production input quantity Q resulting in an output quantity $\tilde{Z} \cdot Q$. As a result, the quantity delivered to the buyer is random, too.

Benchmark

Under centralized or global decision making, i.e. all actions are conducted by one company (indicated by subscript G), the only decision is on the production input quantity Q_G. The profit Π_G can then be formulated in the following way

$$\Pi_G(Q_G) = p \cdot E[\min\{\tilde{Z} \cdot Q_G, D\}] - c \cdot Q_G \quad .$$

From the first order condition (FOC) of the profit function the optimal decision with respect to the production input quantity can be derived as

$$Q_G^* \qquad \text{from} \qquad \int_0^{D/Q_G^*} z \cdot f(z)dz = \frac{c}{p} \quad .$$

For further details please see [4]. Obviously, the SC optimal production input quantity depends on the production cost as well as on the retail price and equals demand inflated by a factor K_G such that

$$Q_G^* = K_G \cdot D$$

$$\text{with} \quad K_G \quad \text{from} \quad \int_0^{1/K_G} z \cdot f(z)dz = \frac{c}{p} \quad \text{and} \quad K_G \geq 1 \quad \text{if} \quad p \geq c/\mu_Z \quad .$$

$p \geq c/\mu_Z$ always holds when profitability of the business is requested, i.e. the expected cost per unit is not larger than the price gained per unit.

A closed-form solution for K_G can be given if the yield rate is uniformly distributed in [0,1] so that $\mu_Z = 0.5$. The respective optimal production input for $p \geq 2 \cdot c$ is then given by

$$Q_G^* = K_G \cdot D \qquad \text{with} \qquad K_G = \sqrt{p/(2 \cdot c)} \geq 1 \quad .$$

Using this result as a benchmark, the situation of local decision making can be evaluated. The Stackelberg game as a game theoretic approach can now be applied with the buyer as leader and the supplier as follower, i.e. the buyer anticipates the supplier's reaction to his own decision.

Supplier Decision

The optimal reaction of the supplier to the buyer's order q maximizes his profit

$$\Pi_S(Q_S|q) = w \cdot E[\min\{\tilde{Z} \cdot Q_S, q\}] - c \cdot Q_S \quad .$$

Thus, from the FOC the optimal supplier decision can be derived as

$$Q_S^* \qquad \text{from} \qquad \int_0^{q/Q_S^*} z \cdot f(z)dz = \frac{c}{w}$$

resulting in $Q_S^*(q) = K_S \cdot q$ and $K_S \geq 1$ if $w \geq c/\mu_Z$.
Similar to the global SC problem, the optimal production input equals the order quantity (here released by the buyer) inflated by some factor K_S. Again, for uniformly distributed yield as above and for profitability of the supplier's business (i.e. $w \geq 2 \cdot c$) the optimal production input is

$$Q_S^*(q) = K_S \cdot q \qquad \text{with} \qquad K_S = \sqrt{w/(2 \cdot c)} \geq 1 \quad .$$

Due to uncertain production processes, the output quantity cannot be determined in advance. This implies that the quantity shipped to the buyer is random, too. This has to be considered when analyzing the buyer's optimal order quantity.

Buyer Decision

From the buyer's perspective, profit is random because of an uncertain delivery quantity $L(Q_S^*(q)\,|q) = E\left[\min\left\{\tilde{Z} \cdot Q_S^*(q), q\right\}\right]$. Anticipating the supplier's reaction, the buyer's profit is given by

$$\Pi_B(q|Q_S^*(q)) = p \cdot E[\min\{L(Q_S^*(q)|q), D\}] - w \cdot E[L(Q_S^*(q)|q)] \quad .$$

From the FOC, and due to $Q_S^*(q) = K_S \cdot q$ the buyer's optimal order quantity is

$$q^* \qquad \text{from} \qquad \int_0^{D/K_S \cdot q^*} z \cdot f(z)dz = \frac{c}{p} + \frac{w}{p} \cdot \frac{1 - F(1/K_S)}{K_S}$$

resulting in $K_S \cdot q^* = K_B \cdot D$ and $1 \leq K_B \leq K_G$

$$\text{with} \qquad K_B \qquad \text{from} \qquad \int_0^{1/K_B} z \cdot f(z)dz = \frac{c}{p} + \frac{w}{p} \cdot \frac{1 - F(1/K_S)}{K_S} \quad .$$

This leads to a reaction of the supplier in form of $Q_S^*(q^*) = K_S \cdot q^* = K_B \cdot D$. Due to $K_B \leq K_G$ it holds that $Q_S^*(q) \leq Q_G^*$. What can further be shown from this analysis is that the optimal production input quantities under global and local decision making differ when $w > c/\mu_Z$ (strict profitability) such that $Q_S^*(q) < Q_G^*$. It follows that SC coordination is not enabled when only a wholesale price is fixed in the parties' contract. This results from the so-called double marginalization effect which states that when both parties aim for positive profits, each SC stage charges a mark-up on the cost it incurs when selling to successive stages. The only case which would result in SC coordination, i.e. $Q_S^*(q) = Q_G^*$, is when the wholesale price equals the expected production cost (i.e. $w = c/\mu_Z$). This scenario, however, violates the business profitability condition for the supplier. For illustration, again, in the case of uniformly distributed yield rate and strict profitability (i.e. $p > w > 2 \cdot c$) we find

$$K_B = K_S \cdot \sqrt{\frac{p}{w \cdot (2 \cdot K_S - 1)}} > 1 \quad .$$

Furthermore, the optimal supplier decision given the buyer's order size is

$$Q_S^*(q^*) = Q_G^* \cdot \sqrt{1/(2 \cdot K_S - 1)} < Q_G^* \quad \text{from} \quad K_S > 1 \quad .$$

2.2 SC Coordination under Random Yield with External Procurement Option

In this section the previous scenario is extended with respect to the opportunity for the supplier to make up for his uncertain production process. Here, we assume that missing units can be procured at a higher price c_E (i.e. $c_E > c/\mu_Z$) from a reliable source if initial production yields less than the amount ordered. The result is that no matter what is yielded from regular production, the order quantity from the buyer will be delivered in full amount. That requires sequential decision making by the supplier. Additionally to deciding on the production input quantity, the supplier has to choose how much to procure in addition once regular production yield is realized.

Benchmark

Again, deterministic demand is considered. Under the centralized setting, the second decision (in a timely matter) in this sequential decision making scenario is on the external procurement quantity M which is $M_G^* = \max\left\{D - z \cdot Q_G^*, 0\right\}$. As a result, demand can be fulfilled completely. The decision prior to the external procurement is obviously the production input quantity. It is chosen in order to maximize the following profit function which accounts for fulfillment of total demand D

$$\Pi_G(Q_G) = p \cdot D - c \cdot Q_G - c_E \cdot E\left[\max\left\{D - \tilde{Z} \cdot Q_G, 0\right\}\right] \quad .$$

The optimal production input quantity which maximizes profit can then be derived from the FOC

$$Q_G^* \qquad \text{from} \qquad \int_0^{D/Q_G^*} z \cdot f(z)dz = \frac{c}{c_E}$$

so that $Q_G^* = K_G \cdot D$ with $K_G > 1$.

The first interesting result from this analysis is that the SC optimal production input quantity is not dependent on the retail price anymore. But still, demand is inflated by a factor which now depends solely on the costs for regular production and external procurement. Having analyzed the global SC problem the result can again be used as a benchmark for the following consideration of local decision makers in a SC.

Supplier Decision

Given the buyer's order quantity, the supplier chooses his production input quantity first and then the external procurement quantity if necessary. In this case, external procurement will be $M_S^* = \max\left\{q - z \cdot Q_S^*(q), 0\right\}$. Prior to procuring externally, regular production takes place. The supplier maximizes his profit which is

$$\Pi_S(Q_S \,|\, q) = w \cdot q - c \cdot Q_S - c_E \cdot E\left[\max\left\{q - \tilde{Z} \cdot Q_S, 0\right\}\right] \quad .$$

The production input quantity then is derived from the FOC

$$Q_S^* \qquad \text{from} \qquad \int_0^{q/Q_S^*} z \cdot f(z)dz = \frac{c}{c_E} \quad .$$

So, we obviously find $Q_S^*(q) = K_G \cdot q$ with the multiplier K_G from the centralized solution. In contrast to the case without external procurement, the optimal production input choice of the supplier is independent of the wholesale price. Analyzing the optimal order quantity of the buyer and comparing the results will then reveal some interesting insights.

Buyer Decision

Due to the fact that the supplier is obliged to procuring missing quantities from regular production, the delivery quantity received by the buyer is not random anymore. This motivates the buyer to order exactly the demand in order to maximize his profit. Now, it can easily be shown that under these circumstances SC coordination will take place. Because the buyer orders exactly the demand, the supplier's production input quantity is $Q_S^*(q) = K_G \cdot D$ which is exactly the SC optimal decision. Accordingly, the decision on external procurement is also identical with the SC optimal one ($M_S^* = M_G^*$). Depending on the parameter setting, profit can be obtained solely by either one of the two actors or be shared among the parties. For $w = c_E$ all profit is left with the buyer whereas the supplier obtains all profit if $w = p$. For

values in between those two extremes, either party receives a share of the profit. Interestingly, in case of stochastic demand, the simple WP contract again fails to achieve coordination of the SC. Under random demand the optimal external procurement quantity results from the newsvendor solution with the order-up-to-level $S_G^* = G^{-1}((p-c_E)/p)$ in the centralized and $S_S^* = G^{-1}((w-c_E)/w)$ in the decentralized setting, where $G(\cdot)$ is the demand's *cdf*. Thus, the optimal quantity to procure externally $(M^* = \max\{S^* - z \cdot Q^*, 0\})$ is different for the centralized and decentralized solution so that double marginalization will occur in the decentralized setting. Thus, even if compensation for insufficient initial production is possible, the simple WP contract will not coordinate the SC when demand is stochastic.

3 Conclusion and Outlook

In this research a supplier-buyer SC was considered where the supplier's production is exposed to random yield so that the buyer faces a stochastic fulfillment level of his order. Putting the respective inventory control problem into a game-theoretic context, it was shown that in case of deterministic end customer demand the WP contract fails to achieve coordination under stochastic yield. If the supplier, however, is able to compensate yield losses via a more costly, but reliable procurement option the analysis reveals that the WP contract is able to coordinate. Interestingly, this coordination property gets lost again if customer demand is stochastic. Thus, it turns out that in the case of yield randomness the coordination power of the WP contract very much depends on the specific SC environment. Future research should contain how the risk of stochastic yield, such as waste of excess units or costs for external procurement can be shared among the SC parties such that coordination is achieved.

References

1. G.P. Cachon. Supply chain coordination with contracts. In A.G. de Kok and S.C. Graves, editors, *Supply Chain Management: Design, Coordination and Operation*, pages 229–339. Elsevier, 2003.
2. M.G. Güler and T. Bilgic. On coordinating an assembly system under random yield and random demand. *European Journal of Operational Research*, 196(1): 342–350, 2009.
3. H. Gurnani and Y. Gerchak. Coordination in decentralized assembly systems with uncertain component yields. *European Journal of Operational Research*, 176(3): 1559–1576, 2007.
4. K. Inderfurth and J. Clemens. The effects of wholesale price contracts for supply chain coordination under stochastic yield. *FEMM Working Paper Series, Faculty of Economics and Management, Otto-von-Guericke University Magdeburg*, 2011.
5. C.A. Yano and H.L. Lee. Lot sizing with random yields: A review. *Operations Research*, 43(2): 311–334, 1995.

Parameters for Production/Inventory Control in the Case of Stochastic Demand and Different Types of Yield Randomness

Karl Inderfurth and Stephanie Vogelgesang

Abstract We consider a single-stage stochastic inventory problem under periodic review and present methods for safety stock determination to compensate quantity uncertainties that are caused by stochastic demand and different types of yield randomness. Taking lead times into account we recommend dynamic safety stocks that vary from period to period. To enable practical manageability we suggest two approaches for calculating static safety stocks that are easy to apply.

1 Introduction

In environments where not only customer demand is stochastic but also production is exposed to stochastic yield, inventory control becomes an extremely challenging task. To cope with the influence of both risks we present two control parameters, which can be used in an MRP-type inventory control system: a safety stock and a yield inflation factor that accounts for yield losses. Considering a single-stage inventory problem under periodic review several authors (see [2, 4]) investigate that the optimal policy for cost minimization results in a critical stock rule in combination with a non-linear order release function which however is difficult to apply in practice. It is well-known that a linear approximation also works quite well in the case of zero production lead time and linear costs for production, stock-keeping, and backlogging (see [1, 5]). Just recently it has been investigated how an effective parameter determination can also be extended to cases with arbitrary lead times (see [3, 6]). Up to now all contributions in this research context refer to production environments that are characterized by stochastic demand and stochastically proportional yield. We will extend the parameter determination approaches to two further well-known types of yield randomness (see [8]), namely binomial and interrupted

Karl Inderfurth, e-mail: karl.inderfurth@ovgu.de
Stephanie Vogelgesang, e-mail: stephanie.vogelgesang@ovgu.de
Faculty of Economics and Management, Otto-von-Guericke University Magdeburg

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_72,

geometric yield. The three mentioned yield models mainly differ in the level of correlation existing for individual unit yields within a single production lot. We show how for all yield models safety stocks can easily be determined following the same theoretical concept. Because in the case of non-zero lead time safety stocks vary from period to period, we additionally present alternative approaches of how these dynamic safety stocks can be transformed into static ones.

2 Linear Control Rule

In the sequel we present a control mechanism which enables us to cope with demand and yield risks and determine appropriate safety stocks (SST).
The following notation is used:

Q_t : released order quantity in period t
CS_t : critical stock for period t
x_t : expected inventory position in period t
SST_t : safety stock for period t
λ : production lead time
$\tilde{Y}(Q)$: random yield (number of good units from a production batch size Q)
$\bar{Y}(Q)$: expected yield ($= E[\tilde{Y}(Q)]$)
$\tilde{Z}$: random yield rate with expectation μ_Z and variance σ_Z^2
$\tilde{D}_t$: random demand in period t with mean μ_D and variance σ_D^2
α : critical ratio (depending on holding and backlogging cost)

Following a critical stock rule with linear order release function, an order Q_t in period t is released if the expected inventory position x_t falls below a critical stock CS_t . If so we order up to CS_t and choose Q_t by multiplying the deviation of critical stock and inventory position with $1/\mu_Z$ to compensate for the expected yield losses. According to that the linear control rule is given by

$$Q_t(x_t) = \max\{(CS_t - x_t)/\mu_z; 0\},$$

where the critical stock is defined as $CS_t = SST_t + (\lambda + 1) \cdot \mu_D$. The expected inventory position at the beginning of period t is calculated by aggregating the net inventory and the yield expectation of all outstanding orders. As the yield risk is considered by adjusting a safety stock, we choose the yield inflation factor to be $1/\mu_Z$, which is different from the approach in [4] and [9] where the yield rate uncertainty is only compensated by modifying this factor. We can determine the safety stock SST_t from

$$Prob\{\tilde{\xi}_t \leq SST_t\} = \alpha, \tag{1}$$

where $\tilde{\xi}_t$ is a random variable defined as total demand deviations minus total yield deviations during the risk period

$$\tilde{\xi}_t = \sum_{i=0}^{\lambda} [\tilde{D}_{t+i} - \mu_D] - \sum_{i=0}^{\lambda-1} [\tilde{Y}(Q_{t-i}) - \bar{Y}(Q_{t-i})] \tag{2}$$

with $E[\tilde{\xi}_t] = 0$ and variance

$$Var\left[\tilde{\xi}_t\right] = (\lambda+1) \cdot \sigma_D^2 + \sum_{i=0}^{\lambda-1} Var\left[\tilde{Y}(Q_{t-i})\right]. \tag{3}$$

Assuming additionally that $\tilde{\xi}_t$ is approximately normally distributed we can solve equation (1) for SST_t resulting in $SST_t = k \cdot \sqrt{Var[\tilde{\xi}_t]}$ with $k = \Phi^{-1}(\alpha)$, where $\Phi(\cdot)$ denotes the standard normal cdf (see [3, 6]).

3 Types of Yield Randomness

In literature (see [8]) three basic types of yield randomness are introduced:

- *Stochastically proportional (= SP) yield*: The production yield $\tilde{Y}(Q)$ from a production batch of size Q is given by $\tilde{Y}(Q) = \tilde{Z} \cdot Q$. The yield rate $\tilde{Z}$ is a random number from interval $[0,1]$ with an arbitrary probability distribution and with mean μ_Z and variance σ_Z . This yield type presumes that yield rate and lotsize are independent.
- *Binomial (= BI) yield*: Binomial yield $\tilde{Y}(Q)$ assumes that the production yield is a random number following a binomial distribution with success parameter p . In this modeling approach the production quality from item to item within a lot is independent of each other and the yield rate parameters are

$$\mu_Z = E[\tilde{Y}(Q)]/Q = p \quad \text{and} \quad \sigma_Z^2 = Var[\tilde{Y}(Q)]/Q^2 = p \cdot (1-p)/Q = \sigma_Z^2(Q).$$

- *Interrupted geometric (= IG) yield*: This modeling approach differs from the other ones because here good items are produced with a success probability p until a failure occurs and all units thereafter are defective. The production yield $\tilde{Y}(Q)$ from a batch of Q units is a random number following an interrupted geometric distribution with probabilities

$$Prob\{Y = k\} = \begin{cases} p^k \cdot (1-p) & , k = 0,1,...,Q-1 \\ p^Q & , k = Q. \end{cases}$$

The resulting yield rate parameters are

$$\mu_Z = \frac{p \cdot (1-p^Q)}{(1-p) \cdot Q} = \mu_Z(Q) \quad \text{and}$$

$$\sigma_Z^2 = \frac{p \cdot (1-p^{1+2Q}) - (1-p) \cdot (1+2Q) \cdot p^{1+Q}}{(1-p)^2 \cdot Q^2} = \sigma_Z^2(Q).$$

4 Parameter Determination Approaches

4.1 Parameter Determination for *SP Yield*

First we apply the *SST* determination procedure to the *stochastically proportional yield* model mentioned above and find by adapting formulae (2) and (3):

$$\tilde{\xi}_t = \sum_{i=0}^{\lambda}[\tilde{D}_{t+i} - \mu_D] - \sum_{i=1}^{\lambda-1}[\tilde{Z}_{t-i} \cdot Q_{t-i} - \mu_Z \cdot Q_{t-i}] - [\tilde{Z}_t \cdot Q_t - \mu_Z \cdot Q_t]$$

$$Var\left[\tilde{\xi}_t\right] = (\lambda+1)\cdot\sigma_D^2 + \sigma_Z^2 \cdot \sum_{i=1}^{\lambda-1} Q_{t-i}^2 + \sigma_Z^2 \cdot Q_t^2 \,.$$

For *SST* calculation in period t the current order quantity Q_t, which still has to be determined, is replaced by the expected order quantity μ_D/μ_Z (inflated expected demand per period) resulting in a dynamic *SST* formula

$$SST_t = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + \sigma_Z^2 \cdot \sum_{i=1}^{\lambda-1} Q_{t-i}^2 + (\sigma_Z^2/\mu_Z^2)\cdot\mu_D^2}\,.$$

Static *SST* approximations might be useful to simplify its application in practice because the dynamic safety stock varies over time due to varying open order quantities Q_{t-i}. We examine two approaches, one which ignores the order variability and another which explicitly considers it. In the first approach all past order quantities Q_{t-i} $(i = 1,...,\lambda-1)$ are replaced by their expected values μ_D/μ_Z leading to

$$SST^{\#1} = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + \max\{\lambda;1\}\cdot(\sigma_Z^2/\mu_Z^2)\cdot\mu_D^2}\,.$$

The second approach is more sophisticated and considers order variability by treating each order as a (a-priori) random variable $\tilde{Q}$ with a total risk depending on demand and yield variability (see [7]). We determine the risk contribution $\tilde{\Delta}$ of a single order in period t as $\tilde{\Delta}_t = (\tilde{Z}_t - \mu_Z)\cdot\tilde{Q}_t$. For a linear control rule with a constant critical stock level, the stochastic order quantity $\tilde{Q}_t$ is generated by $\tilde{Q}_t = (\tilde{D}_{t-1} - \tilde{\Delta}_{t-1})/\mu_Z$. A recursive relationship holds for $\tilde{\Delta}_t$ with $\tilde{\Delta}_t = (\tilde{Z}_t - \mu_Z)\cdot(\tilde{D}_{t-1} - \tilde{\Delta}_{t-1})/\mu_Z$. Because of independence of yield rate from order quantity the risk of an open order in steady-state, where $Var[\tilde{\Delta}_t] = Var[\tilde{\Delta}_{t-1}]$, is given by $Var[\tilde{\Delta}] = \frac{\sigma_Z^2}{\mu_Z^2-\sigma_Z^2}\cdot(\mu_D^2+\sigma_D^2)$. By treating risks from all λ order sizes Q_{t-i} in the same way, the total safety stock results in

$$SST^{\#2} = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + \max\{\lambda;1\}\cdot\frac{\sigma_Z^2}{\mu_Z^2-\sigma_Z^2}\cdot(\mu_D^2+\sigma_D^2)}\,.$$

4.2 Parameter Determination for *BI* Yield

By adapting the same methodology as in the case of *SP* yield we find as dynamic safety stock formula

$$SST_t = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + p\cdot(1-p)\cdot\sum_{i=1}^{\lambda-1} Q_{t-i} + (1-p)\cdot\mu_D}$$

and as static safety stock formula (ignoring order variability)

$$SST^{\#1} = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + \max\{\lambda;1\}\cdot(1-p)\cdot\mu_D}\,.$$

Considering only a myopic order risk contribution $\tilde{\Delta}$ without recursion we get a $\tilde{\Delta}$ variance that in the *BI* yield case also includes a covariance term: $Var[\tilde{\Delta}] = Var[\tilde{Y}(\tilde{D}/p)] + Var[\tilde{D}] - 2\cdot Cov[\tilde{Y}(\tilde{D}/p),\tilde{D}]$. With estimating the covariance appropriately (see [7]) we find the second static *SST* formula

$$SST^{\#2} = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + \max\{\lambda;1\}\cdot\left[(1-p)\cdot\mu_D + 2\cdot\sigma_D^2 - 2\cdot Cov[\tilde{Y},\tilde{D}]\right]}\,.$$

4.3 Parameter Determination for *IG* Yield

In case of *IG* yield the mean yield rate varies with the batch size $\mu_Z = \mu_Z(Q)$, i.e. we have to find an approximation for μ_Z so that we can determine the safety stocks. For a required output in the amount of $\bar{Y}(Q) = \frac{p\cdot(1-p^Q)}{1-p}$ we can calculate the required input

$$Q = \frac{ln(1-\bar{Y}(Q)\cdot(1-p)/p)}{ln(p)}\,,$$

which is only feasible if $\bar{Y}(Q) < p/(1-p)$. With $\mu_Z = \bar{Y}(Q)/Q$ and by replacing the currently required output $\bar{Y}(Q)$ by the mean demand (=mean required output) we get

$$\mu_Z = \frac{\mu_D\cdot ln(p)}{ln(1-\mu_D\cdot(1-p)/p)}\,,$$

which is only feasible if $\mu_D < p/(1-p)$. Following the methodology from above for the *IG* yield model we get the dynamic safety stock formula

$$SST_t = k \cdot \sqrt{(\lambda+1)\cdot\sigma_D^2 + B_1 + C_1}$$

with $B_1 = \frac{1}{(1-p)^2}\sum_{i=1}^{\lambda-1}\left[p\cdot(1-p^{1+2Q_{t-i}}) - (1-p)\cdot(1+2Q_{t-i})\cdot p^{1+Q_{t-i}}\right]$

and $C_1 = \frac{1}{(1-p)^2}\left[p\cdot(1-p^{1+2\cdot\mu_D/\mu_Z}) - (1-p)\cdot(1+2\cdot\mu_D/\mu_Z)\cdot p^{1+\mu_D/\mu_Z}\right]$

as well as the first static safety stock formula (ignoring order variability)

$$SST^{\#1} = k \cdot \sqrt{(\lambda+1) \cdot \sigma_D^2 + \max\{\lambda;1\} \cdot C_1}\,.$$

Calculating the myopic risk for an open order analogously to the approach for *BI* yield, but neglecting the covariance term we find a second static stock formula

$$SST^{\#2} = k \cdot \sqrt{(\lambda+1) \cdot \sigma_D^2 + \max\{\lambda;1\} \cdot \left(Var[\tilde{Y}(\tilde{D}/p)] + \sigma_D^2\right)}$$

where $Var[\tilde{Y}(\tilde{D}/p)]$ is calculated as the variance of an *interrupted geometric* random variable with random trials (see [7]).

5 Conclusion and Outlook

We investigated several different approaches to calculate safety stocks in the case of stochastic demand and random production yield. Thereby, we considered three modeling approaches for yield uncertainty and developed dynamic and static safety stock formulae for each approach. In a comprehensive simulation study we will examine the cost performance of these safety stocks and we will analyze how the cost performance is affected by a misspecification of the yield model.

References

1. S. Bollapragada and T. Morton. Myopic heuristics for the random yield problem. *Operations Research*, 47(5): 713–722, 1999.
2. Y. Gerchak, R. Vickson, and M. Parlar. Periodic review production models with variable yield and uncertain demand. *IEE Transactions*, 20(2): 144–150, 1988.
3. C. Gotzel. *MRP zur Materialplanung für Kreislaufprozesse*. Gabler, 2010.
4. M. Henig and Y. Gerchak. The structure of periodic review policies in the presence of random yields. *Operations Research*, 38(4): 634–643, 1990.
5. W T. Huh and M. Nagarajan. Linear inflation rules for the random yield problem: Analysis and computations. *Operations Research*, 58(1): 244–251, 2010.
6. K. Inderfurth. How to protect against demand and yield risks in MRP systems. *Int. J. Production Economics*, 121(2): 474–481, 2009.
7. K. Inderfurth and S. Vogelgesang. Parameters for production/inventory control in the case of stochastic demand and different types of yield randomness. *FEMM Working Paper Series, Faculty of Economics and Management, Otto-von-Guericke University Magdeburg*, 2011.
8. C.A. Yano and H.L. Lee. Lot sizing with random yields: A review. *Operations Research*, 43(2): 311–334, 1995.
9. P.H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, 2000.

Optimising Procurement Portfolios to Mitigate Risk in Supply Chains

Atilla Yalçın and Achim Koberstein

1 Introduction

In recent years the management of risk in supply chains has become an important issue in the scientific literature. This aspect is of primary interest in market side decisions of a company such as in the procurement of materials. So far, traditional supply chain planning methods solely focus on improving cost efficiency and reducing inventory buffers in supply chains. These approaches are successful as long as the assumption of a stable supply chain environment holds. But when risks on the demand side and on the supply side occur these approaches become contraproductive and make the supply chain more vulnerable. What is needed now are new concepts which improve the flexibility of supply chains even in uncertain environments. In this work we will investigate a mid-term procurement decision where the buyer has to agree with his suppliers on supply contracts while facing demand and supply risk. We assume that the buyer negotiates with multiple suppliers who can supply products with the same quality. The problem we are dealing with is how to design a portfolio of optimal supply contracts in a mid-term planning horizon (e.g. one year) by specifying minimum and maximum quantities of a product in a contract. The objective of our planning problem is to minimise the total expected cost of supply.

The related literature to this work can be divided into four streams. The first stream deals with the question of optimal supply contracts (e.g. [1]). Mainly, the authors use extensions of the newsvendor model and analyse optimal parameter settings of contracts within a supply chain. The second stream of literature is about supplier selection problems where the optimal number of suppliers is determined.

Atilla Yalçın
University of Paderborn, Warburgerstr.100, 33098 Paderborn, Germany, e-mail: yalcin@dsor.de

Achim Koberstein
University of Paderborn, Warburgerstr.100, 33098 Paderborn, Germany, e-mail: koberstein@dsor.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_73,

Jayaraman et al. [3] developed a decision model for the supplier decision problem and also incorporated the allocation of supply into the contracted supplier pool. Harrison et al. [2] analyse the optimal number of suppliers in presence of volume discounts. They explicitly consider in their model the failure of suppliers and use different compensations to include this effect. The third stream of research covers the problem of optimal procurement planning and thereby selecting suppliers in the short-term (e.g. [8]). They consider suppliers where the buyer can order from different discount levels. The question which arises in this setting is to determine the optimal purchasing quantity and time of various suppliers simultaneously. Their approach fits best with models from the lot sizing literature and is suitable for high volume and low value (C type) products. The last stream of literature covers the problem of optimal procurement portfolios with supply options. Martinez de Albeniz and Simchi-Levi [6] first developed a portfolio approach where they combine different supply options for commodity products. In [7] a procurement risk management system is described which was developed at Hewlett-Packard (HP) when they were facing risks on the procurement markets of flash memories.

Our contribution is to provide a new portfolio approach for the mid-term procurement problem to optimise costs and mitigate risks. As opposed to previous portfolio approaches we consider supply and demand risks for multiple periods and address the issue of the supply chain context more precisely.

2 A Portfolio Approach

The fundamentals of Modern Portfolio Theory go back to the work of Markowitz [5]. Markowitz showed that a well selected combination of securities (financial contracts) outperforms every security on its own with respect to return and volatility of an investment. Grounded on this theory today many investors construct their optimal portfolios. Compared to the situation on financial markets the procurement market is not that different. Here, purchasing managers have to make their investments in supply contracts. They have to weigh the cost of the contract against the risk and the flexibility a supply contract provides. Supply disruptions caused by supplier failure or even supplier bankruptcy have to be taken into account as well as uncertainties from the demand side when the available inventory runs out. Especially the problem of supplier failure often leads purchasing managers to choose more than one supplier. Another way to hedge against risks is to hold safety stocks for production but this will result in higher capital costs. Although the management literature uses portfolio techniques to analyse the procurement situation in this case they provide only limited decision support for a profound decision. To get a deeper understanding of supply contracts we will look at the components of a supply contract.

According to [6] we can distinguish between two kinds of contracts: First, fixed (commitment) or long-term contracts are agreements with suppliers where a certain amount of goods is bought in advance. The advantage of these contracts are a higher degree of planning certainty for the suppliers. Hence, suppliers can optimise their

production planning and are able to offer better prices to the buyer. To give the buyer an incentive to buy higher volumes in advance suppliers often give some volume-dependent discounts. Second, short-term contracts are negotiated after an inquiry from the buyer for additional quantities is submitted that goes beyond the scope of the fixed contracts. In order to be prepared for this situation suppliers have to hold some inventory buffer. Another possibility for short-term supply are supplies from non-contracted suppliers or even from the spot market if existent when contracted suppliers cannot provide the requested amounts. Mostly, the price of this short-term transaction can be assumed to be higher than in the case of long-term contracts.

Besides these two kinds of contracts supply options combine aspects of both types of contracts. A supply contract is a right but not an obligation to buy up-to a certain amount of quantities within a pre-defined time to a previously agreed price. It is a powerful means to quantify the supply flexibility needed. Similar to fixed contracts supply options have to be purchased in advance before the actual demand occurs and the suppliers have enough time to considers this in their production planning. The buyer of the supply option has to pre-pay only a fraction of the total price to the supplier up-front, this is called the reservation price. If the buyer requests the reserved amount he has to pay an additional fee, the so-called strike price. All in all the costs of sourcing by supply options tends to be lower than a short-term transaction[1].

If we look at contract agreements in practise they typically have the form of quantity flexible contracts (cf. [1]) where the buyer agrees with the supplier on a minimum and a maximum amount of quantities to purchase. One can express a quantity flexibility contract as a combination of a fixed contract plus a corresponding supply option contract. The minimum quantity can be explained by a fixed or long-term part of this contract. The difference between the minimum and the maximum amount can be considered as the flexible part which can be modeled as supply options and added to the fixed contract. As we can see one can model various kinds of contracts by just specifying the fixed part and the flexible part. The question that appears at this point is whether a portfolio of different contracts from multiple suppliers can be optimised in terms of supply costs and risks from demand and supply side. In the following we strive to formalise this problem by an exact decision model.

3 Model Formulation

In this section we model our portfolio approach for procurement markets as a mathematical programming model. For the sake of simplicity we consider only a procurement problem with one product. We assume that we have a pool of S suppliers where each of the suppliers $s \in \{1..S\}$ submits an offer for the minimum and maximum quantity over the whole planning horizon of T periods. On the basis

[1] From a modeling perspective short-term contracts can also be regarded as a specialised supply option where the buyer does not have to pay a reservation price in advance.

of our previous observation an offer consists of two parts. The first part specifies the pricing of the fixed quantities the buyer will purchase. Depending on the purchased quantities the supplier can offer up-to L discount levels. The second part of the offer specifies additional flexible quantities by supply options. For each amount of supply that the supplier has to reserve the buyer has to pay a reservation price Rp. In case these amounts are called an additional payment of Sp has to be made (strike price). A supplier s can offer V_{sl} different option contracts depending on the chosen discount level $l \in \{1..L\}$ for the fixed part.

In every time period $t \in \{1..T\}$ demand and supply processes take place. For both kinds of processes we assume that they are stochastic and can be expressed as a series of scenarios $r \in \{1..R\}$ where R is the total number of scenarios. We denote by D_t^r the demand in period t that has to be fulfilled in scenario r. Furthermore, we express the supply uncertainty due to supplier failures by a random factor Ω_t^{sr} in each period of a scenario. This value describes the fraction of a purchasing order that is actually delivered. Since we express the uncertainty of the model parameters by scenarios we are able to model our procurement decision problem as a stochastic linear programming (SLP) model. SLP enhances the well-known class of linear programming problems by stochastic elements. For further information see [4].

We employ a two-stage stochastic programming model where at the first stage the decision about the supplier selection and minimum and maximum quantities is taken. As described earlier we allow suppliers to give volume discounts which we model using a mixed integer technique proposed by Stadtler [8]. On the second stage purchasing order allocations are computed for given realisations of the uncertain parameters. This stage serves as an anticipation of a more detailed purchasing order decision in short-term planning in the sense of hierarchical planning. The deterministic equivalent SLP model of our mid-term purchasing planning problem can be stated as follows:

$$\begin{aligned} Min \sum_{s=1}^{S}\sum_{l=1}^{L} F^{sl}\cdot y^{sl} + \sum_{s=1}^{S}\sum_{l=2}^{L} p^{sl}\cdot Q^{s(l-1)}\cdot y^{sl} + \sum_{s=1}^{S}\sum_{l=1}^{L} p^{sl}\cdot q^{sl} + \sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{v\in V_{sl}} Rp^{slv}\cdot o_{max}^{slv} \\ + \sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{v\in V_{sl}}\sum_{r=1}^{R} \pi_r\cdot Sp^{slv}\cdot o^{slvr} + \sum_{t=1}^{T}\sum_{r=1}^{R} \pi_r\cdot h\cdot I_t^r \end{aligned} \tag{1}$$

subject to

$$q^{s1} \le Q^{s1}\cdot y^{s1} \qquad \forall s = 1..S \tag{2}$$

$$q^{sl} \le (Q^{sl} - Q^{s(l-1)})\cdot y^{sl} \qquad \forall s = 1..S, l = 2..L \tag{3}$$

$$\overline{q}^{s} = \sum_{l=2}^{L} Q^{s(l-1)}\cdot y^{sl} + \sum_{l=1}^{L} q^{sl} \qquad \forall s = 1..S \tag{4}$$

$$0 \le o^{slvr} \le o_{max}^{slv} \le M^{slv}\cdot y_{sl} \;\; \forall s = 1..S, l = 1..L, v \in V_{sl}, r = 1..R \tag{5}$$

$$\sum_{t=1}^{T} x_t^{sr} = \overline{q}^s + \sum_{l=1}^{L} \sum_{v \in V_{sl}} o^{slvr} \qquad \forall s = 1..S, r = 1..R \tag{6}$$

$$0 \leq x_t^{sr} \leq Cap_t^s \;\; \forall s = 1..S, r = 1..R, t = 1..T \tag{7}$$

$$I_t^r = I_{t-1}^r - D_t^r + \sum_{s=1}^{S} \Omega_t^{sr} \cdot x_t^{sr} \qquad \forall t = 1..T, r = 1..R \tag{8}$$

$$y^{sl} \in \{0,1\} \;\; \forall l = 1..L \succ SOS1, s = 1..S \tag{9}$$

The objective (1) is to minimise the expected total costs of supply, i.e., of decisions taken in the first and second stage. The binary decision variables y^{sl} represent the decisions about the supplier selection and the appropriate discount level. The subsets of variables associated with each supplier form Special Ordered Sets of Type 1 (SOS1) since only one discount level of a supplier can be valid at a time (Eq. 9). However, a portfolio of more than one supplier is possible.

The first stage costs are displayed in the first line of the objective function. The first term contains fixed costs F^{sl} associated with supplier discount level l ($y^{sl} = 1$, otherwise $y^{sl} = 0$). The next two terms represent the minimum purchase cost at the limits of the discount interval Q^{sl} plus the fixed purchase amounts q^{sl} within the discount intervals both incurring a purchase cost p^{sl} per unit. The last term in the first line expresses costs for additional flexible quantities o_{max}^{slv} caused by the supply option with the reservation price Rp^{slv} .

The second stage costs (the second line) comprise the expected costs when the amount of supply options o^{slvr} are called at the strike price Sp^{slv} in scenario r and the expected holding costs that are charged for the on-hand inventory I_t^r at the end of period t for each scenario r . Here, h denotes the unit holding costs per period and π_r the probability for scenario r .

Inequalities 2 and 3 describe the discount intervals proposed by the suppliers. The total minimum quantity per supplier $\overline{q}^s$ is calculated in Equation 4. The flexibility modeled via supply options is restricted in Equation 5. The actually called quantities o^{slvr} from an option $v \in V_{sl}$ are limited by the supply options o_{max}^{slv} and the maximal available amount M^{slv} . The ordering quantities x_t^{sr} from a supplier s in period t have to be part of the supplier's agreed contract with fixed and flexible quantities $\overline{q}^s$ and o^{slvr} , respectively. In each period t , these quantities must not exceed the capacity boundary Cap_t^s of the supplier (Eq. 7). The inventory level I_t^r is balanced in Equation 8. Note, that supply risk has to be considered at this point and therefore not all ordered quantities can be used to fulfill the demand.

4 Numerical Example & Conclusion

To illustrate the effects of the proposed model we will use a simple numerical example. We consider a procurement situation with $S = 4$ suppliers where each of them offers $L = 3$ discount levels. Our planning horizon for contracts is $T = 50$. To keep this example simple we further assume an independent failure rate ω for all

suppliers leading to random factors $\Omega_t^{sr} \sim B(1, 1-\omega)$ and identical supply costs[2]. The demand D_t^r for each period follows an independent Normal distribution with $D_t^r \sim N(100, \sigma^2)$. Table 1 shows the results of the numerical examples with $R = 30$

	$\sigma = 33$				$\sigma = 66$				$\sigma = 100$			
ω	avg # supl	avg. obj	avg fix qty	avg flex qty	avg # supl	avg. obj	avg fix qty	avg flex qty	avg # supl	avg. obj	avg fix qty	avg flex qty
0	1	238,816	2025	2947	1	244,509	2021	2955	1	251,026	2020	2971
0.01	1	236,525	2000	3137	1	242,448	2000	3303	1	249,393	2004	3485
0.05	1.91	240,025	2039	3370	1.79	245,305	2032	3813	1.73	252,621	2044	4259
0.10	2.69	246,034	2074	3374	2.66	251,649	2065	3860	2.46	258,361	2077	4342

Table 1 Results of the numerical example

scenarios. The values are computed as an average over 100 simulation runs. With an increasing supplier failure rate ω the number of contracted suppliers increases. This is intuitive since the buyer can diversify the contracted suppliers to ensure supply. Interestingly, with an increasing demand volatility the number of suppliers seems to decrease. This observation can be explained by the suppliers' commitment to more flexibility when the fixed volume increases.

In this work we presented a portfolio approach for the mid-term procurement problem, which can be seen as a first step towards a quantitative evaluation of procurement strategies when risk hedging mechanisms have to be considered explicitly.

References

1. R. Anupindi and Y. Bassok. Supply contracts with quantity commitments and stochastic demand. In S. Tayur, M. Magazine, and R. Ganeshan, editors, *Quantitative Models for Supply Chain Management*, pages 199–232. Kluwer Academic Publsihers, 1999.
2. G. Harrison, S. Moritz, and R. Pibernik. The optimal number of suppliers in the presence of volume discounts and different compensation potentials – an analytical and numerical analysis. *European Business School Research Paper Series*, 09-03, 2008.
3. V. Jayaraman, R. Srivastava, and W.C.Benton. Supplier selection and order quantity allocation: A comprehensive model. *The Journal of Supply Chain Management*, Spring:35(2): 50–59, 1999.
4. P. Kall and S.W. Wallace. *Stochastic Programming*. John Wiley & Sons, 1994.
5. Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1): 77–91, 1952.
6. V. Martinez-de-Albeniz and D. Simchi-Levi. A portfolio approach to procurement contracts. *Production and Operations Management*, 14(1): 90–114, 2005.
7. V. Nagali, J. Hwang, D. Sanghera, M. Gaskins, M. Pridgen, and T. Thurston et al. Procurement risk management (PRM) at Hewlett-Packard company. *Interface*, 38(1): 61–75, 2008.
8. H. Stadtler. A general quantity discount and supplier selection mixed integer programming model. *OR Spectrum*, 29(4): 723–745, 2007.

[2] $p^{s1} = 100, p^{s1} = 98, p^{s3} = 95$ with $Q^{s1} = 500, Q^{s2} = 1000, Q^{s3} = 5000$ and only one supply option ($\|V_{sl}\| = 1$) for each price level $Rp^{s1} = 21, Rp^{s2} = 14, Rp^{s3} = 12$ with $Sp^{s1} = 92, Sp^{s2} = 92, Sp^{s3} = 92$ and flexibility limits $M^{s1} = 0, M^{s2} = 1000, M^{s3} = 5000$

Using Simulation for Setting Terms of Performance Based Contracts

James Ferguson and Manbir Sodhi

Abstract This paper discusses the use of a simulation model for setting terms of performance based contracts. Metrics commonly used in measuring service and supply performance are examined for their utility in achieving the outcomes of interest for contracts, and the correlations amongst these metrics are determined using simulation models. These correlations are used to highlight the most significant measurable quantities, and the performance limits for these are specified so as to achieve desired outcomes for various system performance indicators. The impact of penalties and incentives are also evaluated using the simulations. Practical issues such as obsolescence, reliability, and cost are also discussed. The exploration of this concept in setting terms for a contract at the Naval Undersea Warfare Center is also illustrated.

1 Introduction

The model discussed in the paper simulates an organization's inventory that holds spare subassemblies for a fleet of end-products used by the organization in carrying out a specific mission. This environment was created to mimic the US Navy Torpedo Enterprise. In the Torpedo Enterprise, torpedoes are routinely exercised and deployed within the larger US Navy Submarine Enterprise. Due to continual exercising and deployment, torpedoes are constantly being serviced and tested before being returned to the fleet for further exercises and fleet use. When the torpedoes are serviced, their subassemblies are tested. If the subassemblies are found to be defective, they are replaced from the inventory. Prior to 1986, these spare subassemblies

James Ferguson
Code 81: Torpedoes, Naval Undersea Warfare Center, Newport, RI 02882 e-mail: james.c.ferguson@navy.mil

Manbir Sodhi
Dept. of Mech, Industrial & Systems Eng., Univ. or Rhode Island, Kingston, RI 02881, USA, e-mail: sodhi@uri.edu

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 465
DOI 10.1007/978-3-642-20009-0_74,

were procured by the US Navy Supply system using fully documented disclosure packages at low risk. However, starting in 1986 with the Packard Commission Report [7], "A quest for excellence", and continuing into 2002, a number of acquisition reform initiatives were issued that changed the way the US Navy and other organizations acquired new systems. [8] started to shift the focus of the acquisition world from processes to outcomes. Currently, when a contract for a torpedo subassembly is constructed, it is set up with an initial base year of production and optional production years in the future. Minimum and maximum procurement levels and lot size pricing is also included in the contract. The idea is that after the base year, the Torpedo Enterprise will determine the need for more subassemblies. This need is based on failure and usage rates, among other factors. Much time and energy is spent in determining these amounts. In recent years, there have been efforts within the Department of Defense (DoD) to implement performance-based logistics contracting methods. However, [3] noted that there has been reluctance by managers to implement the performance-based logistics (PBL) construct within the DoD. In [3] it is stated that PBL has received a poor reputation because of the unconventional techniques it employs. Nevertheless, [3] provides figures showing recent cost and time savings within DoD, which can be directly attributed to a program's use of PBL strategies despite these unconventional techniques. [6] also recognize the difficulties encountered when seeking to implement PBL contracts. [6] also provides guidance with respect to what type of contract (fixed-price, cost-plus, or PBL) should be used in certain contractual situations.

2 The Model

The model creates a simulation of a PBL contract. PBL seeks to minimize the number of subassemblies kept in the customer's inventory by having the contractor restock the inventory as the subassemblies are used. The contractor obligates himself to support the subassembly to a specified operational availability while keeping no more than a prescribed number of subassemblies at the customer's facility. In this PBL environment, the customer/contractor interface is at the shelf; literally, where the subassemblies are stored at the customer's facility. The customer provides the contractor with up-to-date inventory levels, which trigger the contractor's decision to restock. For instance, if a torpedo subassembly is found to be defective during testing, a spare would be removed from the inventory to replace it. When this removal from inventory occurs, the contractor is notified that the inventory level has decreased by one. At this point, the contractor will produce the needed subassembly and restock the customer's inventory. Decisions regarding the production and storage of the subassemblies are the contractor's prerogative. The contractor could make one large batch and store the subassemblies themselves or create a dedicated production line that produces the subassemblies as needed. The contract for this environment should be as simplistic as possible. The contractor is required to maintain a given inventory level in order to achieve a specified operational availability at the

customer's facility. This inventory level will be based upon the customer's usage rate, which will be provided to the contractors bidding on the contract.

The model is a newsvendor type model with several variations. The classic newsvendor problem is a single period mathematical model used to determine optimal inventory levels when the demand is uncertain [9]. The newsvendor problem has been used as a starting point for analyzing many scenarios. A review of some extensions can be found in [5]. Among the cases that can be related to the analysis of contractor performance are [2], [1], [6] and others. In [2] a newsvendor model is used to structure a scenario when a single newsvendor is served by several suppliers, some or all of whom may be unreliable. In [1], a vendor commits to an initial purchase; and a demand estimate is subsequently revealed. [6] evaluates PBL as a strategy for purchasing the "results of a product" as opposed to buying the actual repair parts, spares and maintenance activities. In [4], it is noted that PBL specifies outcomes, not numbers of spare parts or hours of maintenance. The emphasis of the contract is on metrics to be achieved by the contractor, (in [4] the metrics are operational availability and readiness risk) not the way in which the contractor must achieve the specified metrics.

The goal of the model created for the torpedo enterprise and presented in this paper is to identify correlations between metrics relating to subassemblies storage and the number of subassemblies kept in the customer's inventory. The metrics that are strongly correlated to the subassembly's inventory level can then be used in a competitive contracting environment to discriminate between bids. The model first generates 2000 subassemblies with randomly selected values for the following subassembly attributes; operational tempo (OPTEMPO), failure rate, logistical delay time, minimum operational availability, shortage cost and storage cost. OPTEMPO is the expected usage rate of the products being supported by the subassembly inventory for a given time period. In the case of the Torpedo Enterprise, this would be the number of torpedoes expected to be received for maintenance, cleaning, testing and reassembly. The failure rate is the expected failure rate, based on exercise runs, of the subassembly used within the end product. Logistical delay time is specified as the time between removing a subassembly from the shelf and a replacement subassembly arriving from the contractor. The operational availability is equal to 100% minus the failure rate of a subassembly for a given period. Shortage cost is the cost per day of not having a usable subassembly when one is needed. This could be considered the cost of a worker's downtime (the work that could have been done), or the cost of penalties due to delays. Storage cost is the cost of storing one subassembly for one day at the customer's facility. These metrics were given high and low values. Each product was randomly assigned either the high or low value for each of its attributes. The high and low values for these metrics can be viewed below.

Once the 2000 product's attributes are randomly assigned they are individually placed in the simulation. The simulation first inputs the product's attributes and simulates a subassembly's life-cycle over 1000 periods (in this case days). Next, the simulation randomly generates failures throughout the time periods, which correspond to the subassembly's failure rate and OPTEMPO. Defective subassemblies

Table 1 Parameter Values for Model

METRIC	High Value	Low Value
OpTempo	1000/year	500/year
Failure Rate	10%	1%
Logistical Delay Time	28 days	7 days
Min. Operational Availability	99%	95%
Shortage Cost	$ 1000	$ 500
Storage Cost	$100	$ 50

are replaced in the inventory depending on the logistical delay time (e.g. if the logistical delay time is 28 days and a defective subassembly is replaced on day 100 a replacement will arrive on day 128). The operational availability is calculated over the 1000 days by calculating the total demand minus the total unfilled demand divided by the total demand. For example, if 100 parts fail (demand) and 5 were unable to be filled because of an inventory level of zero, the operational availability would be (100 - 5) / 100 = 95%. Finally, the subassembly's inventory level is increased incrementally until the actual operational availability equals or exceeds the minimum acceptable operational availability. This inventory level is then stored and the model simulates the life-cycle of the next subassembly in the same way. The total cost per period is also calculated by multiplying a positive inventory level by the storage cost per period, or the absolute value of a negative inventory level (unfilled demand) by the shortage cost per period.

When all of the subassemblies' life-cycles have been simulated and the needed inventory levels have been determined, the next step in the model is to calculate the correlations between the needed inventory levels and the OPTEMPO, failure rate, logistical delay time, minimum acceptable operational availability, storage cost and shortage cost. This correlation is determined by creating a graph with the inventory level on the y-axis and the metric in question on the x-axis. A linear trend line is added to the graph and its R-squared value is calculated to determine the fit between the trend line and the data.

3 Results

After running the model, the correlations were observed to determine if subassembly inventory levels could be used as a predictor for any of the other six metrics. The model showed that there was virtually no correlation between the subassembly inventory levels and either shortage or storage costs. The model also showed that the correlations between subassembly inventory levels and the OPTEMPO as well as the minimum operational availability were also negligibly low. However, model

showed an obvious positive correlation between subassembly inventory levels and the subassembly failure rate, and there was a slight positive correlation between the subassembly inventory levels and logistical delay time. Because the subassembly inventory level and subassembly failure rate are positively correlated, if Contractor A proposed keeping a larger subassembly inventory at the customer's facility than Contractor B, it could be surmised that Contractor B's subassemblies would have lower failure rates than Contractor A's.

4 Implementations

There are several implementations for this model when setting terms for a performance-based logistics type contract. These implementations include helping to specify meaningful metrics for assessing the contractor's performance, creating fair rewards and penalties within the contract, and differentiating between contractor bids in a competitive environment.

Operational availability was used in the model as the main metric in determining contractor performance. This is due to that fact that in the PBL environment the customer is concerned, primarily, with subassembly availability in inventory and reliability in the field. Also, when assessing a contractor's performance, it is important to use metrics that will provide the customer with insight into the contractor's supply chain efficiency and manufacturing processes. If a subassembly's operational availability begins to fluctuate or drop, it is clear that something affecting the contractor's process reliability has changed. Perhaps the subassembly is failing more often due to manufacturing error, or the logistical delay time is increasing because of an issue with the contractor's supply chain. If these metrics are also tracked, the cause for the fluctuation can be further isolated. The metric used to track contactor performance should be sensitive enough to indicate to the customer when a corrective action might need to be taken in order to maintain the needed inventory levels and therefore the needed operational availability. [10] provides an in-depth look at many performance metrics and was a valuable resource for identifying and defining metrics for this model.

The model can be used to specify realistic and fair rewards and penalties within the contract. "What-if" scenarios can be played out using expected shortage and storage costs to observe the effects of increased failure rates or logistical delay times. Metrics can be adjusted to see how the changes would affect contractor performance on the whole, and provide a decision tool for setting values for the rewards and penalties.

Lastly, the correlation between subassembly inventory levels and failure rates can be used when differentiating between bids in a competitive environment. In the request for proposal (RFP), the customer can require the prospective contractors to provide the subassembly inventory level they would need to keep at the customer's facility in order to attain the specified minimum operational availability.

As was discussed previously, and therefore subassembly failure rates are correlated to the needed subassembly inventory level, inferences can be made about the product quality a contractor is expecting to deliver. For example, it could be concluded that a contractor proposing to keep an inventory level of 5 subassemblies would deliver a higher quality product than a contractor proposing an inventory level of 10 subassemblies.

5 Conclusion

In conclusion, the model developed during this project simulates a PBL support environment. This model is meant to mimic a support environment, which could be utilized within the US Navy's Torpedo Enterprise. This model showed that a correlation exists between certain performance metrics associated with subassemblies, namely logistical delay time and failure rate. These correlations can then be leveraged during contract negotiations in a competitive contracting environment.

References

1. A. Bensoussan, Q. Feng, and S. P. Sethi. A Two-Stage Newsvendor Problem with a Service Constraint. *School of Management, The University of Texas at Dallas, Richardson, TX.*, 2004.
2. M. Dada, N. Petruzzi, and L. B. Schwarz. A Newsvendor's Procurement Problem When Suppliers are Unreliable. *OPERATIONSMANAGEMENT*, 9(1): 9–32, 2007.
3. A. Fowler. Misunderstood superheroes. *Defense AT&L*, Jan-Feb 2009.
4. K. Kang, K. H. Doerr, and S. Sanchez. A Design of Experiments Approach to Readiness Risk Analysis. In *PROCEEDINGSOF THE 2006 WINTER SIMULATION CONFERENCE*, 2006.
5. M. Khouja. The Single-PeriodNews-Vendor Problem: LiteratureReview and Suggestions for Future Research. *OMEGA*, 27: 537–553, 1999.
6. S. Kim, M. A. Cohen, and S. Netessine. Performance Contracting in After-Sales Service Supply Chains. *MANAGEMENT SCIENCE*, 53(12): 1843–1858, 2007.
7. D. Packard. A Quest for Excellence: Final Report to the President by the President's Blue Ribbon Commission on Defense Management. *Washington: Government Printing Office*, 1986.
8. W. J. Perry. Acquisition reform, a mandate for change. Testimony before a Joint Hearing of the Senate Committee on Armed Services and Senate Committee on Governmental Affairs. *Washington, DC: Government Printing Office*, 1994.
9. E. Porteus. *Stochastic Inventory Theory, HANDBOOKS IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE,VOL. 2: STOCHASTIC MODELS.* Heyman, D.P. and Sobel, M.J. Elsevier Science Publishers, Amsterdam, 1991.
10. Defense Acquisition University. Performance measure definitions. *https://acc.dau.mil/CommunityBrowser.aspx?id=22646*, March 2010.

A Credit Risk Modelling Approach to Assess Supplier Default Risk

Stephan M. Wagner and Christoph Bode

Abstract The purpose of this paper is to quantify the supplier default risk in a buying firm's supplier portfolio. Based on credit risk models, we develop a methodology that buying firms can use to pro-actively determine their exposure to supplier default risk. To illustrate the proposed methodology, we use empirical data pertaining to supplier portfolios of executive-size car models from three German automotive OEMs. We show that some supplier portfolios are exposed to higher default risk which places them at a disadvantage, because they face a higher probability that the supply of components can be disrupted and cars cannot be built and sold.

1 Purpose

To date, the assessment and quantification of supplier default risk that goes beyond the analysis of supplier firm credit ratings (e.g., Dun & Bradstreet) has remained largely unexplored. Reasons are that the prediction of the behaviour of firms that are connected in a supply network is complex and that supplier defaults are correlated and not independent events [6]. Furthermore, firms often do not have access to data necessary to perform comprehensive supplier default risk assessments. In consequence, there is a lack of quantitative models for the systematic quantification of supplier default risk. Therefore, the first goal of this paper is to develop a method based on portfolio credit risk models that can be used to assess supplier default risk. Consistent with recent research's call that buying firms need to manage the risk in their supplier *portfolio* (as opposed to managing the risk of individual suppliers)

Prof. Dr. Stephan M. Wagner
Swiss Federal Institute of Technology Zurich (ETH Zurich), Chair of Logistics Management, Scheuchzerstrasse 7, 8092 Zurich, e-mail: stwagner@ethz.ch

Dr. Christoph Bode
Swiss Federal Institute of Technology Zurich (ETH Zurich), Chair of Logistics Management, Scheuchzerstrasse 7, 8092 Zurich, e-mail: cbode@ethz.ch

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 471
DOI 10.1007/978-3-642-20009-0_75,

[7], the proposed method can be applied on the portfolio level. The second goal is to exemplify the application of the method by analyzing real-life supplier portfolios and assessing and comparing the default risk inherent in these portfolios.

2 Research Approach

In finance and the actuarial sciences, various models for estimating the distribution of possible credit losses from a portfolio have been developed. Drawing on these literatures, we apply the Bernoulli model and the CreditRisk+ framework to supplier portfolios to estimate supplier default risk. These models require three basic input parameters: (1) Default probability and default probability volatilities, (2) exposure at default, and (3) default correlations.

2.1 Essential Risk Attributes of Counterparties

First, the *default probability* (DP) of a firm represents the probability that the firm goes bankrupt during a certain time period, typically one year [3]. Basically, two approaches can be used to estimate the DP of a supplier firm. The first approach uses publicly available market or firm data. For instance, the well-known *Z-score* [1] uses financial ratios (e.g., liquidity ratio, leverage ratio) to predict a firm's bankruptcy. The second approach uses ratings from rating agencies such as Moody's, Standard & Poor's, or Fitch. A rating agency determines the credit worthiness of a firm and assigns a rating which is expressed in a letter system such as AAA or B–. The rating from a letter system can be easily transformed to a DP by using historic default frequencies, for instance Moody's historic bond default frequencies. In sum, in the first approach, collecting market data of the firm and modelling the DP has to be done by the researcher, whereas in the second approach, the analysis of the firm is done by rating agencies. Obviously, the advantage of the second approach is that it requires less effort for data collection and analysis. One disadvantage of ratings, however, is that rating agencies only rate a limited number of firms. For example, in total only 59 automotive suppliers were rated by Moody's in 2009 [5]. Furthermore, another disadvantage of ratings is that they are not transparent and may not be reliable.

Second, in banking, the *exposure at default* (EAD) of an obligor contains two parts: drawn ($OUTST$) and undrawn ($COMM$) loan at the time before default, where $EAD = OUTST + \gamma COMM$ and γ is the expected portion of the commitments likely to be drawn prior to default. Generally, "exposure at supplier default" is measured in monetary units. In purchasing, there are only few situations where EAD can be estimated in terms of monetary units. In case of common raw materials, for example, a buying firm can switch from a contracted supplier to a spot market to satisfy its demand. Given a spot market, the exposure at contract breach by a sup-

plier is the difference between the spot market unit price and the contract unit price multiplied by the volume of the contract [4]. However, in the automotive industry, parts, components, modules, or systems are highly customer-specific, so that spot markets with price information do not exist. Therefore, the "exposure at supplier default" is difficult to estimate accurately. Besides the availability of such information, another difficulty is that supplier defaults are rare events and empirical reports of the loss from defaults are often not available. Nevertheless, one can identify determinants for defaults (e.g., complexity of the components, availability of suppliers, ease of switching suppliers) and derive quantitative measures for exposure (on non-monetary scales), and use this information in the credit risk model. Then, expected loss of single default can be calculated in the following way. The buying firm assigns to a supplier i a default probability (DP_i), an exposure at default (EAD_i), and a loss fraction called the "loss given default" (LGD_i) which describes the fraction of the exposure subject to be lost in the considered time period [3]. Then, the loss $\widetilde{L}_i$ in case of the default of supplier i is:

$$\widetilde{L}_i = EAD_i \times LGD_i \times L_i \text{ with } L_i = 1_D, P(D) = DP_i \tag{1}$$

where D denotes the event that the supplier defaults in a certain period of time (most often one year). L_i is a Bernoulli random variable and $P(D)$ denotes the probability of D. The expectation of any Bernoulli random variable 1_D is its event probability. The expected loss (EL_i) of supplier i as the expectation of its corresponding loss variable $\widetilde{L}_i$ is determined by:

$$EL_i = E[\widetilde{L}_i] = EAD_i \times LGD_i \times P(D) = EAD_i \times LGD_i \times DP_i \tag{2}$$

Third, similar to the correlation between obligors in a loan portfolio, recent studies have shown that the defaults of suppliers in a portfolio can also be correlated [2, 6]. *Default correlation* between firms in credit risk modelling should be the same from a bank and from an automotive OEM point of view. There are well established explanations about correlation in the credit risk modelling research, for instance the state of the overall economy or the situation in the particular industry. Correlation is arguably the most challenging part of credit risk modelling. One basic idea is to treat default probabilities as random variables. The default frequency of companies in the same rating class can vary from year to year. In the Bernoulli model the default correlations are fully captured by the covariance structure of the stochastic default probabilities. In the CreditRisk+ model the correlation is introduced by randomization of default intensity and sector analysis [8].

2.2 Empirical Setting

We opted for the German automotive industry as the setting for our empirical analysis. On the one hand, automotive OEMs are large and powerful customers, on the other hand, suppliers are critical for automotive OEMs to achieve and sustain com-

petitive advantage due to the high degree of outsourcing and the innovation that comes from the suppliers. The high criticality of suppliers for the OEM's success coupled with frequent supplier defaults observed in the industry warrants an investigation of supplier default risks which can inform industry practice and research.

The database "Who Supplies Whom" (published by SupplierBusiness) was used to obtain comprehensive information about current car models, components, and modules as well as suppliers. Based on this information, we constructed and analyzed supplier portfolios for three executive-size car models which have similar target customers and belong to a similar price range: the *BMW 5-series* (platform: E60; SOP: 2003), the *Audi A6* (platform: C6; SOP: 2004), and the *Mercedes E-class* (platform: W211; SOP: 2002).

Company ratings and default probabilities were obtained from the AMADEUS database (published by Bureau van Dijk). AMADEUS focuses on European companies and provides standardized annual reports, financial ratios, and information on business activities and ownership structures on approximately 11 million firms. AMADEUS also offers a rating, the *MORE rating*, a credit risk product from ModeFinance. The MORE rating is used for obtaining default probabilities.

As discussed earlier, exposure is difficult to measure in the automotive industry. Therefore, we have chosen switching costs as a proxy for exposure. While exposure measures the value subject to loss at default, switching costs measure the cost of switching away from a supplier after a default. Both concepts are part, component, module, system, and supplier market specific.

Estimating a dollar value of switching cost of a certain component is almost impossible for a broad range of components. Therefore, we aimed at assessing how difficult it is to switch away from a supplier of a certain component. Based on a standardized questionnaire, two expert informants, who were familiar with the automotive supply markets, were asked to estimate the switching cost for each component on a five-point (Likert-type) rating scale (anchored at "1: very low" and "5: very high"), which we denoted as switching cost rating (SCR). Thus, "switching costs" are measured on an arbitrary unit. In case a single supplier delivered several components, the switching cost ratings of the individual components were aggregated, because the unit of analysis is the individual supplier, not a component. For any given supplier that supplies n components, we allowed for three different possibilities to calculate the overall SCR from the single SCR_j s of the n delivered components:

I $SCR = \sum_{j=1}^{n} SCR_j$ (sum)
II $SCR = \frac{1}{n} \sum_{j=1}^{n} SCR_j$ (mean)
III $SCR = \max_{j \in \{1,\ldots,n\}} \{SCR_j\}$ (maximum)

Scenario I assumes that the components' SCR_j are linear and additive. Thus, the more components a supplier delivers, the higher its total SCR. The caveat of this approach is that, if a supplier delivers five easily replaceable components, each with $SCR_j = 1$, the total SCR will be 5, which is equal to a supplier that delivers a single component with $SCR_j = 5$ (i.e., very hard to substitute). This may or may not be realistic. Also, following scenario I a supplier can receive a total SCR which

exceeds 5. Therefore, scenario II takes the average SCR of all components in the supply scope as the rating of the supplier. Now, if a supplier supplies 5 components with each $SCR_j = 1$, the supplier is assigned a total $SCR = 1$. Finally, scenario III assumes that the most difficult component to replace in the supply scope is decisive for the supplier's SCR. For instance, if one supplier delivers two components u and v with $SCR_u = 4$ and $SCR_v = 3$, the supplier is assigned $SCR = 4$. Similar to scenario II, the supplier's SCR ranges between 1 and 5.

3 Findings

We analyze the supplier portfolios for the BMW 5-series, the Audi A6, and the Mercedes E-class, and demonstrate how the Bernoulli model and the CreditRisk+ framework can be applied to these three portfolios. The results provide various information which support the automotive OEMs in supply chain risk quantifications and management from various aspects. The data shows that the three portfolios share numerous common suppliers. Specifically, we find the highest expected loss within the Mercedes E-class portfolio. For instance, for scenario I, the 99-percentile loss level of the BMW 5-series portfolio is 69, the Audi A6 is 94, and the Mercedes E-class is 96.

4 Managerial Implications

In this paper, credit risk models are introduced to the supply chain context to quantify supplier default risk. Specifically, the Bernoulli model and the CreditRisk+ model were examined in terms of their applicability. For buying firms, credit risk models can provide an integral framework for analyzing default probabilities, exposure of suppliers, and the default correlation structure among suppliers. First, the central message from the conducted analyses is that knowing the suppliers is key. Automotive OEMs should pay attention to the risk attributes of their suppliers. The data collection process itself enables the buying firm to know its own supplier portfolio better in a systematic way. Based on the collected data, the OEM may keep a watch list of suppliers which have high default probability in the next year. The OEM may also analyze the suppliers with very high exposure and identify them as strategic partner or consider the development of alternative suppliers.

Second, the outputs of the models may support various decision making processes in supply chain risk management. Using the default event distribution and the loss distribution of the supplier portfolio, the OEMs can plan its human and capital resources more efficiently.

For instance, if the analysis shows that the default risk of the portfolio is high in the next year, it is perhaps advocated to invest more resources in managing the supplier base, enhancing supplier development, or switching away from a supplier with high default probability. The portfolio loss distribution supports buying firms in creating risk mitigation strategies (e.g. insurances).

5 Conclusions

Advanced credit risk models can be adapted to assess supplier default risk. As the application of credit risk models requires a large amount of information on the suppliers, the buying firms (including OEMs) are in the best position to do the analysis. The analysis process may become a tool for OEMs to quantify and manage supplier default risk. The models can also be applied to other industries where the buying firm is concerned about potential supplier defaults.

References

1. Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4): 589–609, 1968.
2. Volodymyr Babich, Apostolos N. Burnetas, and Peter H. Ritchken. Competition and diversification effects in supply chains with supplier default risk. *Manufacturing & Service Operations Management*, 9(2): 123–146, 2007.
3. Christian Bluhm, Ludger Overbeck, and Christoph Wagner. *An Introduction to Credit Risk Modeling*. Chapman & Hall, Boca Raton, FL, 2002.
4. Cagri Haksöz and Ashay Kadam. Supply portfolio risk. *The Journal of Operational Risk*, 4(1): 59–77, 2009.
5. Timothy L. Harrod. *Global auto supplier industry*. Moody's Investors Service, New York, NY, 2009.
6. Stephan M. Wagner, Christoph Bode, and Philipp Koziol. Supplier default dependencies: Empirical evidence from the automotive industry. *European Journal of Operational Research*, 199(1): 150–161, 2009.
7. Stephan M. Wagner and Jean L. Johnson. Configuring and managing strategic supplier portfolios. *Industrial Marketing Management*, 33(8): 717–730, 2004.
8. Tom Wilde. *CreditRisk+ A credit risk management framework*. Credit Suisse First Boston, London, UK, 1997.

Considering Distribution Logistics in Production Sequencing: Problem Definition and Solution Algorithm

Christian Schwede and Bernd Hellingrath

1 Introduction

The aim of production planning in the automotive industry is to terminate the production cycle for every order. Because modern OEM use mixed-model assembly lines (MMAL) a central task within this planning is the order sequencing (OS). It transforms an unsorted group of orders of a certain time period into an optimal sequence.

Since the automotive industry is production dominated, the OS is mainly based on production related goals. Nevertheless, the sequence has a strong impact on performance of distribution logistics (DL) as well, which has the task to transport finished vehicles to logistic centres and dealers. Hence, in this paper we analyse if requirements from DL are already considered and if not can be considered in future OS.

The paper is structured as follows. We start identifying requirements from production and DL that concern the OS, sum up results from a literature review in Section 2 and discover a shortcoming in the consideration of DL related requirements. Thus, in Section 4 we present the problem class Distribution-oriented Car Sequencing (DCS) as an extension of traditional Car Sequencing (CS). In Section 4, a solution algorithm for the new problem is outlined. Section 5 shows first results of a real world scenario, summarises the results and presents ideas for further research.

Christian Schwede
Fraunhofer IML, Joseph-von-Fraunhofer-Str. 2-4, 44227 Dortmund, Germany, e-mail: Christian.Schwede@iml.fraunhofer.de

Bernd Hellingrath
Information Systems and Supply Chain Management, University of Münster, Leonardo-Campus 3, 48149 Münster, Germany, e-mail: Bernd.Hellingrath@wi.uni-muenster.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_76,

2 Requirements Concerning Order Sequencing and Literature Review

In this section we identify impacts of the order sequence on production and DL and requirements that can be derived from these impacts.

Production is the primary driver for sequencing. Especially cost and throughput time efficiency of the final assembly is affected by the sequence. The requirements are minimisation of work overload by avoiding time intensive jobs in series and minimisation of idle time by ensuring an even workload if possible.

The sequence also has a strong impact on DL. The time that passes between the production end of cars with same directions influences stock levels, delivery times and reliability as well as capacity utilisation of the transports. Grouping order with the same destination considering logistic batch sizes [4, p.6] due to departure dates would increase the distribution performance. To achieve this, a mapping of cars to distribution transports with departure dates must be considered as an input for OS.

There are three problem classes in literature that deal with the sequencing problem: Mixed-Model Sequencing (MMS), CS and Level Scheduling (LS) [1]. The literature review was done by classifying 38 papers that use these three classes due to their focus on supply logistics, production and DL. Concerning the available volume in the publication we can only outline the results of the review here. Most of the papers reviewed deal with production, one-third consider supply logistics and the minority - only three - address distribution logistics, which can be approved by earlier investigations [5]. Among the three, only two deal with reaching given departure dates [7, 10] but without considering transports, which means that stock levels cannot be calculated since a car that misses its transport cannot be assigned to the next one when there is no information on destinations or transports. [5] take transports into account, but only on a daily basis, so that neither capacity utilisation of single transports nor concrete departure dates of e.g. scheduled transports are considered. The potential of considering DL stated earlier as well as the insufficient scientific coverage of this topic so far, motivates our work of considering DL in OS comprehensively. We begin with the definition of a new problem class in Section 4.

3 Distribution-Oriented Car Sequencing

In this section, a new distribution-oriented problem class for the sequencing problem will be presented. We chose traditional CS as a basis for the new class for three reasons. First requirements regarding data acquisition are lower than with MMS [3, p.2], which has a higher level of detail. Second the added complexity integrating DL appears to be manageable. Third analyses indicate that using LS to minimise overload of workstations cannot compete with CS [2]. Following the notation used by [1] summarised in table 2 we present the traditional CS first. Column 2 states if the measure is an input (I) or variable (V).

Table 1 Notation of Car Sequencing Problem according to [1]

$t = 1,\ldots,T$	I	Production cycles
$m \in M$	I	Car models (cars/orders with similar properties)
$o \in O$	I	Options of a car
$a_{m,o}$	I	Model m has option o
$x_{m,t}$	V	Model m is produced in cycle t
$d_m \in \mathbb{N}$	I	Demand of model m
$H_o : N_o$	I	Sequencing constraint of option o (max. H_o orders with option o in subsequence of length N_o)
$z_{o,t} \in \{0,1\}$	V	Constraint of option o is violated

The aim is to minimise equation (3)

$$\sum_{o \in O} \sum_{t=1}^{T} z_{o,t} \tag{1}$$

subject to constraints (2)-(11).

$$x_{m,t} \in \{0,1\} \qquad \forall m \in M; t = 1,\ldots,T \tag{2}$$

$$\sum_{m \in M} x_{m,t} = 1 \qquad t = 1,\ldots,T \tag{3}$$

$$\sum_{t=1}^{T} x_{m,t} = d_m \qquad \forall m \in M \tag{4}$$

$$\sum_{m \in M} \sum_{t'=t}^{\min(t+N_o-1,T)} a_m, o \cdot x_{m,t'}$$

$$- \left(1 - \sum_{m \in M} a_m, o \cdot x_{m,t}\right) \cdot BI \leq H_o + BI \cdot z_{o,t} \qquad \forall o \in O; t = 1,\ldots,T \tag{5}$$

Constraint (2) and (5) assure that every cycle contains exactly one model. Constraint (8) states that all the demanded models are produced and constraint (11) ensures that $z_{o,t}$ count the number of violations, with BI being a big integer value.

Now we will extend traditional CS by distribution transports for finished cars. Table 2 shows the additional notation of the DCS followed by the Problem definition.

Table 2 Notation of Distribution-oriented Car Sequencing Problem

$\vartheta = 1,\ldots,\Theta$	I	Distribution transports
$dep_\vartheta \in \{1,\ldots,T\}$	I	Departure cycle of transport ϑ; $dep_0 = 0$ and $dep_{\Theta+1} = T$
$cap_{\vartheta,m} \in \mathbb{N}_0$	I	Number of model m planned for transport ϑ
$alloc_{\vartheta,t} \in \{0,1\}$	V	Model that is produced in cycle t is allocated to transport ϑ
$stocktime_t \in \{0,\ldots,T\}$	V	Cycles, the model produced in cycle t waits for transportation
$extra_t \in \{0,1\}$	V	Extra transports is needed for the model produced in cycle t

Minimise $\sum_{o \in O} \sum_{t=1}^{T} z_{o,t}$, $\sum_{t \in T} stocktime_t$ and $\sum_{t \in T} extra_t$ subject to constraints (2)-(11) and (6)-(11).

$$alloc_{\vartheta,t} \in \{0,1\} \qquad \vartheta = 1,\ldots,\Theta; t = 1,\ldots,T \tag{6}$$

$$alloc_{\vartheta,t} = \Phi_{\vartheta,t} \cdot \left(1 - \min\left(1, \sum_{\vartheta'=1}^{\vartheta-1} \Phi_{\vartheta',t} \cdot \Psi_{\vartheta',t}\right)\right) \cdot \Psi_{\vartheta,t} \qquad \vartheta = 1,\ldots,\Theta; t = 1,\ldots,T \tag{7}$$

$$\Phi_{\vartheta,t} = \min\left(1, \max\left(0, dep_\vartheta - t + 1\right)\right) \qquad \vartheta = 1,\ldots,\Theta; t = 1,\ldots,T \tag{8}$$

$$\Psi_{\vartheta,t} = \sum_{m \in M} \left(\min\left(1, cap_{\vartheta,m} \cdot x_{m,t}\right) \cdot \min\left(1, \max\left(0, cap_{\vartheta,m} - \sum_{\tilde{\vartheta}=1}^{\max(1,\vartheta-1)} \max\left(0, \sum_{t'=dep_{\tilde{\vartheta}-1}+1}^{dep_{\tilde{\vartheta}}} x_{m,t'} - cap_{\tilde{\vartheta},m}\right) + \sum_{t'=dep_{\vartheta-1}+1}^{t-1} x_{m,t'}\right)\right)\right) \qquad \vartheta = 1,\ldots,\Theta; t = 1,\ldots,T \tag{9}$$

$$stocktime_t = \sum_{\vartheta=1}^{\Theta} alloc_{\vartheta,t} \cdot (dep_\vartheta - t) \qquad t = 1,\ldots,T \tag{10}$$

$$extra_t = 1 - \sum_{\vartheta=1}^{\Theta} alloc_{\vartheta,t} \qquad t = 1,\ldots,T \tag{11}$$

Every of the Θ distributions transports has a latest production end cycle dep_ϑ that a corresponding model has to meet to reach the transport. An assignment plan from models to transports $cap_{\vartheta,m}$ has already been done in advance and has to be met by the solution. The actual assignment $alloc_{\vartheta,t}$ states that the model finished in cycle t is transported by transport ϑ. This assignment is done scanning the sequence from left to right, connecting the current model with the next transport that has free capacity. Constraint (6) - (9) ensure a correct setting of $alloc_{\vartheta,t}$ with $\Phi_{\vartheta,t}$ and $\Psi_{\vartheta,t}$ being auxiliary variables. $\Phi_{\vartheta,t}$ is 1 if transport ϑ departs in cycle t or later. $\Psi_{\vartheta,t}$ is one if the model of cycle t can be assigned to transport ϑ due to capacity availability. Constraint (7) is 1 if model of cycle t fits on transport ϑ and is not already assigned to an earlier transport. Furthermore, the new problem is extended by two objectives. The overall stock time that a model waits for transportation is to be minimised ($stocktime_t$) as well as the amount of extra transports needed for models missing their transport ($extra_t$).Constraints (10) and (11) sum up the values for the new objectives.

4 An Algorithm to Solve DCS

Since classic CS is NP-hard [6], meta heuristics have been widely applied in real-case scenarios, while exact optimisation methods performed poorly in comparison [9]. Therefore we designed an Evolutionary Algorithm (EA) combined with a Local Search (LS) to solve the DCS. The LS follows a steepest descent approach with swap, shift and k-swap operators, while mutation operators of the EA are used to diversify the solution. The three fitness dimensions are weighted and added to obtain the overall fitness. To minimise computing time for the fitness evaluation we distribute the overall fitness on the single orders, which allows us to adapt only the values necessary after every operation. Furthermore, the sequencing constraints of the orders are stored in bit vectors allowing a high efficient matching operation. For more details on the algorithms we refer to [8].

5 Scenario-Based Evaluation and Conclusion

To evaluate the algorithm we use a real-case scenario from a German car manufacturer. The data was assembled in the European research project InTerTrans (see www.in-ter-trans.eu). For validation, a simulation model of the site was created using OTD-NET (see www.otd-net.de). A plant with two production lines (average daily capacity of 500 cars each) with 12 and 13 rules respectively is used. For first investigations connections to 5 distribution centres (scheduled, unscheduled and combined transportation) were selected to be optimised within the OS. Results from the simulation are evaluated for different transport concepts. First, concerning scheduled train transportation it has been shown that using DCS capacity utilisation has been increased from 76% to 100%, reducing the additional truck transports by 34%. Second, the stock of finished cars at the port has been reduced by $27,5\%$. Third, for the truck transportation on demand a relevant KPI is the time needed to fill a transport with a capacity of 9 cars. This time has been reduced by 90%. Concerning the sequencing rules the total number of violations has slightly increased in comparison to the productive system. For the first line by $0,7\%$ and for the second line by $1,8\%$. These minor changes are acceptable since they were mainly caused by violations of rules that already had a high violation level.

In this paper we started describing requirements for OS concerning production and DL. Then we summed up results of a review based on these requirements. DCS was defined as a new problem class and a solution algorithm outlined. The presented first results show a high potential for DL performance optimisation using DCS. Further work includes the extension of the model with up to 34 distribution centres as well as the consideration of the supply processes.

References

1. Nils Boysen, Malte Fliedner, and Armin Scholl. Sequencing mixed-model assembly lines: Survey, classification and model critique. *European Journal of Operational Research*, 192(2): 349–373, 2009.
2. Nils Boysen and Karl-Werner Hansmann. *Variantenfließfertigung: Univ., FB Wirtschaftswiss., Diss.–Hamburg, 2005*, volume 49 of *Betriebswirtschaftliche Forschung zur Unternehmensführung*. Dt. Univ.-Verl., Wiesbaden, 1. Aufl., 2005.
3. Uli Golle, Nils Boysen, and Franz Rothlauf. Analysis and Design of Sequencing Rules for Car Sequencing. Mainz, 2009.
4. A. Hermes, M. Preuss, A. Wagenitz, and B. Hellingrath. Integrierte Produktions- und Transportplanung in der Automobilindustrie zur Steigerung der ökologischen Effizienz. *Tagungsband 14. Magdeburger Logistiktagung – 'Sustainable Logistics'*, pages 183–195, 2009.
5. Mingzhou Jin, Yi Luo, and D. Sandra Eksioglu. Integration of production sequencing and outbound logistics in the automotive industry. *International Journal of Production Economics*, 113(2): 766–774, 2008.
6. Tamás Kis. On the complexity of the car sequencing problem. *Operations Research Letters*, 32(4): 331–335, 2004.
7. H. Robin Lovgren and J. Michael Racer. Algorithms for mixed-model sequencing with due date restrictions. *European Journal of Operational Research*, 120(2): 408–422, 2000.
8. Christian Schwede, Katja Klingebiel, Thomas Pauli, and Axel Wagenitz. Simulationsgestützte Optimierung für die distributionsorientierte Auftragsreihenfolgeplanung in der Automobilindustrie. In L. März, W. Krug, O. Rose, and G. Weigert, editors, *Simulation und Optimierung in Produktion und Logistik*, VDI-Buch. Springer, 2010.
9. Christine Solnon, Dat van Cung, Alain Nguyen, and Christian Artigues. The car sequencing problem: Overview of state-of-the-art methods and industrial case-study of the ROADEF'2005 challenge problem. *European Journal of Operational Research*, 191(3): 912–927, 2008.
10. Yuanhui Zhang, B. Peter Luh, Kiyoshi Yoneda, Toshiyuki Kano, and Yuji Kyoya. Mixed-Model Assembly Line Scheduling Using the Lagrangian Relaxation Technique. *IIE Trans. Schedul. Logist*, 32: 125–134, 2000.

II.7 Scheduling and Project Management

Chair: Prof. Dr. Rainer Kolisch (Technische Universität München)

Scheduling addresses the question of allocating resources over time in order to perform a set of operations subject to specific objectives. Project Management incorporates the short-term question of scheduling the operations of one or multiple projects as well as the long-term question such as the valuation of projects and the determination of project portfolios. We invite submission in both areas.

Hybrid Flow Shop Scheduling: Heuristic Solutions and LP-Based Lower Bounds

Verena Gondek

Abstract This paper is concerned with hybrid flow shop scheduling taking account of different additional constraints, especially transportation requests. Our objective is total weighted completion time minimization. We develop a fast (two phase) heuristic solution technique, which is based on simple dispatching rules. To evaluate our approach empirically, we introduce and analyze lower bounds based on the LP relaxation of time-indexed mixed-integer formulations.

1 Introduction

This research is motivated by a real-life problem arising in steel producing industries. For monitoring the production process, samples of steel, slack, and raw-iron are periodically analyzed in an automatic laboratory. The organization of the workflow in this laboratory can be classified as a dynamic hybrid flow shop (HFS) scheduling problem with transportation requests, no-waiting-time constraints as well as blocking constraints and jobs arriving over time. The objective is to minimize total weighted completion time. Due to vast restrictions in computational time, we develop a fast heuristic solution technique, which is presented in Section 2. The used LP-based lower bounds are analyzed in Section 3. A brief overview on our computational results and some conclusions are given in Section 4.

2 A Two-Phase Heuristic Solution Method

The method introduced in this section breaks the HFS problem with transportation into two subproblems, which are solved sequentially. In a first step, we allocate the samples to the available machines and determine the corresponding job sequence

Verena Gondek
University of Duisburg-Essen, Department of Mathematics, e-mail: verena.gondek@uni-due.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_77,

for each machine type. We refer to this subproblem, which can be classified as a $HFS|no-wait|\sum w_jC_j$ -problem, as the *sequencing problem*. Based on the solution of the sequencing problem, we determine a schedule for the fleet of robots to handle the transportation tasks in a second step. This *routing problem* can be characterized as a $Pm|sds,prec,delay|\sum w_jC_j$ -problem, according to the common three-field classification scheme. For instance, Carlier et al. (2010) use a similar decomposition.

2.1 Step 1: Solution of the Sequencing Problem

We are convinced that the adaptation of suitable dispatching rules in combination with list scheduling is the only possibility to comply with the present runtime requirements. There are several results in the literature showing that simple dispatching rules can be useful within hybrid flow shop scheduling (e.g., Azizoglu et al. (2001) or Kyparisis and Koulamas (2001) utilize the well-known SPT/WSPT rule to tackle the problems $FFm||\sum C_j$ and $Fm||\sum w_jC_j$). Therefore, we develop a hybrid algorithm, which exploits the advantages of different fast dispatching rules. As pointed out by Queyranne and Schulz (2006), only a job-driven list scheduling approach, instead of Graham's non-idling, is appropriate to deal with a job-related objective function like total weighted completion time. In the following, we first introduce our dynamic list scheduling approach, and afterwards we present the different methods that can be used to compute initial job sequences.

Based on a given job sequence, we schedule the jobs according to the following greedy list scheduling strategy: Select the first job in the list and insert it into the current partial schedule such that its processing on the first machine can be started as early as possible. In case of no-waiting-time constraints, the start times on the other stages follow implicitly. Nevertheless, the strategy can be easily adapted to a problem with unlimited intermediate buffer.

There are often considerable differences in the workload of the production stages, i.e., there may be one or more bottleneck stages that have a great impact on the performance of the whole system. In the following, we take account of this fact, and according to the paper of Azizoğlu et al. we define a bottleneck stage i^* as the stage with maximum relative workload (i.e., the total workload divided by the number of available machines). Thus, we start with the following two initial job sequences.

1. The jobs are sequenced in increasing order according to their weighted total processing time (i.e., the sum of the processing times of a job over all stages in relation to its weight).
2. The jobs are sequenced in increasing order according to their weighted processing time on the bottleneck stage i^* .

The second sequence neglects all available information about the workload of the remaining stages, hence, we consider three additional sequencing methods. To that

end, we approximatively solve the problem $Pm|r_{i^*j}, pmnt|\sum w_j C_j$ using the preemptive version of the WSPT rule for the bottleneck stage i^*. The release dates are defined as $r_{i^*j} := \sum_{h=1}^{i^*} p_{(h-1)j}$, with $p_{0j} := r_j$ for all $j \in J$. Let C_j^* denote the completion time of job j in the schedule generated by preemptive WSPT. We then consider the sequence $j_1, j_2, \ldots, j_n$, if:

3. $C_{j_1}^* \leq C_{j_2}^* \leq \ldots \leq C_{j_n}^*$.
4. $M_{j_1}^* \leq M_{j_2}^* \leq \ldots \leq M_{j_n}^*$, with $M_j^* := C_j^* - \frac{1}{2} p_{i^*j}$.
5. $C_{j_1}^*(\frac{1}{2}) \leq C_{j_2}^*(\frac{1}{2}) \leq \ldots \leq C_{j_n}^*(\frac{1}{2})$, where $C_j^*(\frac{1}{2})$ is the point in time when half of the processing requirement p_{i^*j} is completed.

The list scheduling approach is applied to all of the five sequences then and the one with the best objective function value is chosen. The method does not specify which machine of a certain type should be used to process a job. If we assume that all machines on one stage are identical and are located at the same position, we can allocate the jobs to the machines arbitrarily.

2.2 Step 2: Solution of the Routing Problem

In the second step, we have to solve a $Pm|sds, prec, delay|\sum w_j C_j$-problem, whereby the robots are defined as identical machines in parallel. Based on the solution of the sequencing problem obtained in Step 1, we construct a precedence graph for the transportation tasks that have to be performed. A sequence-dependent setup is induced by an empty travel between two consecutive transportation tasks, which are performed by the same robot. Finally, the delays result from the processing of a job on a machine that must be completed before it can be surveyed to the next machine.

To the best of our knowledge, there is no attempt in the literature to treat precedence constraints and sequence-dependent setup times in a parallel machine environment simultaneously, except for a paper published by Hurink and Knust (2001), who focus on makespan minimization and problem complexity. Nevertheless, our heuristic solution method is motivated by a paper of Weng et al. (2001), who exclusively deal with sequence-dependent setup times in a parallel machine environment and total weighted completion time objective.

In each step of the heuristic, we determine a list of available jobs, according to the precedence graph. Dependent on the current partial schedule, we compute the earliest point in time $f_k(r)$ when an available transportation task k could be started on robot r. For each of those pairs (k,r) we calculate a priority index $\kappa_{rk} := f_k(r) + \frac{s_{kl_r} + p_k}{w_k}$, where l_r is the last job that was loaded onto robot r and therefore, s_{kl_r} denotes the arising setup time under usage of robot r. p_k defines the transportation time of job k and w_k is the corresponding weight. We choose the pair (k,r) with the smallest index κ_{rk} then, and insert it into the partial schedule.

2.3 Dynamization

We are concerned with a practical application where jobs arrive over time, thus, all information about a job is available as soon as it reaches the laboratory. The heuristic solution technique described above is very fast and therefore, it is possible to re-optimize the whole system, every time a new job arrives. Naturally, jobs, which are already in process, cannot be considered in a re-optimization step.

3 LP-Based Lower Bounds

To evaluate the solutions determined with the algorithm presented above, we use two lower bounds provided by the LP relaxations of time-indexed MIP formulations. In the first one, we minimize $\sum w_j C_{mj}$ subject to

$$\sum_{j=1}^{n} y_{ijt} \leq n_i \qquad \text{for } i \in I, t = 0,\ldots,T \tag{1}$$

$$\sum_{i=1}^{m} y_{ijt} \leq 1 \qquad \text{for } j \in J, t = 0,\ldots,T \tag{2}$$

$$\sum_{t=0}^{T} y_{ijt} = p_{ij} \qquad \text{for } i \in I,\ j = 1,\ldots,n \tag{3}$$

$$C_{ij} = \frac{p_{ij}}{2} + \frac{1}{p_{ij}} \sum_{t=0}^{T} \left(t + \frac{1}{2}\right) y_{ijt} \qquad \text{for } j \in J \tag{4}$$

The binary variable y_{ijt} equals 1, if job j is processed on a machine of type i in the time interval $[t, t+1]$, and 0 otherwise. Thus, y_{ijt} equals 0, if $t < r_j + \sum_{l=1}^{i-1} p_{lj}$ or $t \geq T - \sum_{l=i+1}^{m} p_{lj}$. T can be any feasible upper bound on the planning horizon. In the second formulation, we minimize the same objective subject to

$$\sum_{t=0}^{T} x_{ijt} = 1 \qquad \text{for } i \in I,\ j \in J \tag{5}$$

$$\sum_{j=1}^{n} \sum_{s=t-p_{ij}+1}^{t} x_{ijs} \leq n_i \qquad \text{for } i \in I, t = 0,\ldots,T \tag{6}$$

$$\sum_{i=1}^{m} \sum_{s=t-p_{ij}+1}^{t} x_{ijs} \leq 1 \qquad \text{for } j \in J, t = 0,\ldots,T \tag{7}$$

$$C_{ij} = \sum_{t=0}^{T} t x_{ijt} + p_{ij} \qquad \text{for } i \in I,\ j \in J \tag{8}$$

Here, the binary variable x_{ijt} equals 1, if job j is started on a machine of type i at time t, and 0 otherwise. Hence, x_{ijt} equals 0, if $t < r_j + \sum_{l=1}^{i-1} p_{lj}$

or $t > T - \sum_{l=i}^{m} p_{lj}$. In both formulations, we additionally consider the following constraints to model the processing order among the stages.

$$C_{ij} \overset{(=)}{\leq} C_{(i+1)j} - p_{(i+1)j} \qquad \text{for } j \in J,\ i = 1\ldots,m-1 \tag{9}$$

$$r_j \leq C_{1j} - p_{1j} \qquad \text{for } j \in J \tag{10}$$

It can be verified that the second formulation dominates the first one, using the substitution $y_{ijt} = \sum_{t-p_{ij}<s\leq t} x_{ijs}$ (due to the proof of Dyer and Wolsey (1990) for the corresponding single machine problem). Furthermore, it can be shown that the LP relaxations of both formulations dominate the completion time based model introduced by Schulz (1996). To proof this, we use a result of Goemans (1996), who established the equivalence of the first time-indexed formulation with a completion time based formulation in the case of a single machine.

To consider transportation requests, we model the robots as an additional production stage, which has to be visited several times; i.e., if we have an instance with m stages, the model with transportation contains $m+1$ stages and $2m+1$ production steps for each job, respectively.

4 Computational Results and Conclusions

To evaluate the performance of our algorithm as well as the strength of the used bounds, we randomly generated a set of test instances consisting of 5 up to 30 jobs and 2 or 3 production stages as well as 2 or 3 robots. For each combination we generated 10 instances. All of used the parameters were drawn as integers from a uniform distribution, with $r_j \in [0,50]$ (Set 1) and $r_j \in [0,100]$ (Set 2), respectively, and $p_{ij} \in [20,40]$. The processing times in the real-life situation are notably larger than the transportation times, hence, we chose the latter from the interval $[1,5]$.

The results are summarized in Table 1. The average relative error (in %) in comparison to the lower bounds is given for each dimension. The computation of LB2 is very time-consuming for larger instances, therefore, we just give the values for the small instances here. The values in parantheses in the last row show the overall average deviation for the small instances in case of LB1, to provide an opportunity to compare the performance of both bounds.

Altogether, the results are promising. This is also true in case of our real-life problem, where we achieved a cost reduction of more than 40% in comparison to the current control system. The results show that LB2 dominates LB1 in practice, too. Nevertheless, the mean deviation is not that big (cf. Table 1), consequently, the computational effort to compute LB2 might not be justified in terms of its purpose within this paper.

Table 1 Relative error vs. lower bounds

	Set 1				**Set 2**			
	2 Robots		3 Robots		2 Robots		3 Robots	
$n \times m$	LB1	LB2	LB1	LB2	LB1	LB2	LB1	LB2
5×2	6.36	4.91	4.43	3.03	6.62	5.35	5.02	3.76
5×3	11.25	9.34	9.13	7.27	6.18	5.20	3.89	2.94
10×2	13.03	10.93	8.51	6.50	12.67	10.88	6.96	5.24
10×3	18.32	17.00	10.88	9.65	15.96	14.03	10.92	9.08
15×2	18.62	16.93	9.73	8.18	21.88	19.87	12.58	10.71
15×3	20.80	19.42	12.73	11.43	22.14	20.60	12.06	10.65
20×2	22.20	-	12.48	-	27.62	-	14.77	-
20×3	25.46	-	14.79	-	25.95	-	17.17	-
30×2	20.64	-	12.48	-	23.66	-	12.09	-
30×3	29.41	-	17.57	-	25.38	-	17.56	-
Average	**18.61** (14.73)	**13.09**	**11.11** (9.24)	**7.68**	**18.81** (14.24)	**12.65**	**11.30** (8.57)	**7.06**

References

1. M. Azizoğlu, E. Çakmak, and S. Kondakcı. A flexible flowshop problem with total flow time minimization. *Eur. J. Oper. Res.*, 132(3): 528–538, 2001.
2. J. Carlier, M. Haouari, M. Kharbeche, and A. Moukrim. An optimization-based heuristic for the robotic cell problem. *Eur. J. Oper. Res.*, 202: 636–645, 2010.
3. M.E. Dyer and L.A. Wolsey. Formulating the single machine sequencing problem with release dates as a mixed integer program. *Discrete Appl. Math.*, 26: 255–270, 1990.
4. M.X. Goemans. A supermodular relaxation for scheduling with release dates. In *Queyranne (Eds.), Integer Programming and Combinatorial Optimization, Lect. Notes Comput. Sci.*, pages 288–300. Springer, 1996.
5. J. Hurink and S. Knust. List scheduling in a parallel machine environment with precedence constraints and setup times. *Oper. Res. Lett.*, 29: 231–239, 2001.
6. G.J. Kyparisis and C. Koulamas. A note on weighted completion time minimization in a flexible flow shop. *Oper. Res. Lett.*, 29: 5–11, 2001.
7. M. Queyranne and A.S. Schulz. Approximation bounds for a general class of precedence constrained parallel machine scheduling problems. *SIAM J. Comp.*, 35(5): 1241–1253, 2006.
8. A.S. Schulz. Scheduling to minimize total weighted completion time: Performance guarantees of LP-based heuristics and lower bounds. *Lect. Notes Comput. Sci.*, 1084: 301–315, 1996.
9. M.X. Weng, J. Lu, and H. Ren. Unrelated parallel machine scheduling with setup consideration and a total weighted completion time objective. *Int. J. Prod. Eco.*, 70: 215–226, 2001.

Solving the Earth Observing Satellite Constellation Scheduling Problem by Branch-and-Price

Pei Wang and Gerhard Reinelt

1 Introduction

An important economical issue for a space agency like China Center for Resource Satellite Data and Applications (CRESDA) is the optimization of the schedule of its earth observing satellites. The management of the constellation consists of selecting a feasible subset of requests and scheduling the observation activities of the satellites to acquire images as well as scheduling the download activities to transmit the images back to a set of ground stations, with the objective to maximize the summed rewards of the targets observed and image-downloaded. These selected and scheduled sequence of observations and downloads must comply with visibility time-windows, observing and downloading durations, satellite memory capacity, and minimum transition times between observations and downloads.

Most of the literature in the field of satellite scheduling considered single-satellite single-orbit problems. Moreover, few researchers took into account both observation scheduling and download scheduling, instead studied these two parts separately. So far as we know, the integrated scheduling problem of observations and downloads for a satellite constellation received rare attention. Only Frank [3] used a constraint-based interval planning framework to model the problem and proposed a stochastic heuristic search method, but they did not give any experimental result. Florio [2] and Bianchessi [1] addressed the problem with a look-ahead insert heuristic and a FIFO heuristic respectively. Yet they did not provide the mathematical model of the problem and neither of them considered the station contention between two different satellites when downloading consecutively.

Pei Wang
Institute of Computer Science, University of Heidelberg, Germany e-mail: pei.wang@informatik.uni-heidelberg.de

Gerhard Reinelt
Institute of Computer Science, University of Heidelberg, Germany e-mail: gerhard.reinelt@informatik.uni-heidelberg.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 491
DOI 10.1007/978-3-642-20009-0_78,

In this paper we propose a constraint-based formulation for the satellite constellation scheduling problem considering the download contentions. The original model is decomposed into a set packing master problem and a series of longest path sub-problems with resource constraints and time windows. The linear-relaxed master problem is solved by CPLEX and the sub-problem is solved by a labeling-based dynamic programming. Given a fractional master problem solution, a branching scheme on original model's flow variables is also discussed.

2 Problem Formulation

The satellite constellation scheduling problem can be characterized as follows:

- $N = \{1,...,n\}$: set of targets; $w_i > 0, i \in N$: reward of servicing target i. $P = \{1,...,n\}$ and $D = \{n+1,...,2n\}$ denote the observations and downloads of all the targets respectively.
- $J^0 = P \cup D \cup \{0\} = J \cup \{0\}$: job set, where 0 is a dummy job that represents the starting and ending of the job sequence on every satellite.
- K : set of satellites; $K_i, i \in J$: subset of satellites on which job i can be serviced; $c_k > 0, k \in K$: satellite storage capacity (the unit is second, which is corresponding to the observing duration maximally allowed without downloading); $s_{ijk}, i \neq j \in J^0, k \in K$: minimum transition time between servicing two jobs consecutively; $d_{ik}, i \in J, k \in K_i$: requested servicing duration for job i by satellite k. Note that $d_{ik} = d_{(i-n)k}, i \in D$.
- $OTW = \{[a_{ikq}, b_{ikq}] | i \in N, k \in K_i, 1 \leq q \leq cn_{ik}\}$: set of observation time windows, where a_{ikq} and b_{ikq} are the earliest and latest start time for satellite k to observe target i at time window q, and cn_{ik} is the number of time windows between i and k.
- $DTW = \{[a_{kp}, b_{kp}] | k \in K, 1 \leq p \leq nb_k\}$: set of download time windows, where a_{kp} and b_{kp} are the earliest start time and latest end time for satellite k to download to ground stations at time window p, and nb_k is the number of time windows between k and stations. W : index set of all the download time windows for all the satellites, overlapping ones from different satellites are only counted once.
- 0, T : scheduling start and end time respectively.

We have four types of variables:

- binary variable x_{ikq} and y_{ikp} : indicating whether the observation and image-download of target i are adopted within the time window $[a_{ikq}, b_{ikq}]$ and $[a_{kp}, b_{kp}]$ respectively. Another binary variable y'_{kp} is also introduced, indicating whether download time window $p \in W$ is used for satellite k.
- binary variable z_{ijk} : indicating whether job j is directly processed after job i by satellite k.
- integer variable t_i : start time of job i.
- integer variable l_{kr} : indicating the memory level at time r for satellite k.

A constraint-based model is as follows:

$$\max \sum_{i\in N} w_i \sum_{k\in K_i} \sum_{1\le q\le cn_{ik}} x_{ikq} \tag{1}$$

$$\text{s.t.} \sum_{k\in K_i} \sum_{1\le q\le cn_{ik}} x_{ikq} = \sum_{k\in K_i} \sum_{1\le p\le nb_k} y_{ikp} \le 1, \quad \forall\, i \in N \tag{2}$$

$$\sum_{k\in K} y'_{kp} \le 1, \quad \forall\, p \in W \tag{3}$$

$$x_{ikq} = 1 \Rightarrow a_{ikq} \le t_i \le b_{ikq}, \quad \forall\, i \in N, k \in K_i, 1 \le q \le cn_{ik} \tag{4}$$

$$y_{ikp} = 1 \Rightarrow a_{kp} \le t_{i+n} \le b_{kp} - d_{ik}, \quad \forall\, i \in N, k \in K_i, 1 \le p \le nb_k \tag{5}$$

$$t_i < t_{i+n}, \quad \forall\, i \in P \tag{6}$$

$$z_{ijk} = 1 \Rightarrow t_i + d_{ik} + s_{ijk} \le t_j, \quad \forall\, i \ne j \in J, k \in K \tag{7}$$

$$x_{ikq} = 1, i \in P, 1 \le q \le cn_{ik} \Rightarrow l_{ka_{ikq}} + d_{ik} = l_{k(b_{ikq}+d_{ik})}, \quad \forall\, k \in K \tag{8}$$

$$y_{(i-n)kp} = 1, i \in D, 1 \le p \le nb_k \Rightarrow l_{ka_{kp}} - d_{ik} = l_{kb_{kp}}, \quad \forall\, k \in K \tag{9}$$

$$l_{k0} = l_{kT} = 0, \quad \forall\, k \in K \tag{10}$$

The objective function (1) is to maximize the total rewards of serviced targets. Constraint (2) regulates that the observation or download of every target is executed at most once. (3) constrains that at most one satellite is downloading images to stations at any given time. (4) and (5) bound the start times of adopted observations and downloads. (6) prescribes the sequence between the observation and the download of the same target. (7) restricts the transition time between consecutive activities. (8), (9) and (10) together describe the satellite memory level change when taking activities, and the initial and end condition of the memory level.

3 Algorithm

3.1 Set Packing Reformulation

To decompose the original formulation, we define the following parameters:

- Λ_k : set of all the partial schedules for a single satellite.
- binary parameter ρ^k_{ju} : indicating whether schedule $u \in \Lambda_k$ services target $j, j \in N$.
- binary parameter σ^k_{pu} : indicating whether schedule $u \in \Lambda_k$ adopts download time window $p \in W$ for satellite k .
- ω^k_u : rewards of schedule $u \in \Lambda_k$.

- binary variable q_u^k : indicating whether u is adopted as the final schedule for satellite k .

The original problem is then formulated as the following set packing problem:

$$\text{(SP)} \qquad \max \sum_{k\in K}\sum_{u\in\Lambda_k} \omega_u^k q_u^k \tag{11}$$

$$\text{s.t.} \sum_{k\in K}\sum_{u\in\Lambda_k} \rho_{ju}^k q_u^k \le 1, \quad \forall\, j \in N \tag{12}$$

$$\sum_{k\in K}\sum_{u\in\Lambda_k} \sigma_{pu}^k q_u^k \le 1, \quad \forall\, p \in W \tag{13}$$

$$\sum_{u\in\Lambda_k} q_u^k \le 1, \quad \forall\, k \in K \tag{14}$$

The sub-problem of searching for the partial schedule with the largest reduced cost of SP is as follows:

$$max\{\sum_{j\in N}(w_j-\lambda_j)\rho_{ju}^k - \sum_{p\in W}\tau_p\sigma_{pu}^k - \theta_k | u\in\Lambda_k, k\in K\} \tag{15}$$

$$\text{s.t.} \qquad \text{(4) to (10)}$$

where λ_j , τ_p and θ_k are dual variables for (12), (13) and (14) respectively.

3.2 Dynamic Programming for Sub-Problems

The sub-problem is to determine the partial schedule with the largest reduced cost regulated by (15), which corresponds to finding a longest path in a graph with time windows and resource constraints (denoted as LPSTRC). The following graph is constructed:

$G=(V,A)$, $V=OTW_k\cup DTW_k\cup\{\alpha,\beta\}$, where $OTW_k, DTW_k, \alpha, \beta$ are observation time windows, download time windows for satellite k , source node and sink node respectively. For $[a_{ikq},b_{ikq}],[a_{jkp},b_{jkp}]\in V$, if $a_{ikq}+d_{ik}+s_{ijk}\le b_{jkp}$, then there exists an arc from $[a_{ikq},b_{ikq}]$ to $[a_{jkp},b_{jkp}]$ in A . The arc definition between observations and downloads or between different downloads respects the same time compatibility as above. In order to tackle the memory capacity, we attach a series of labels to each node in V , and every label L_v has five parts, i.e., (T,M,RC,U,S) , representing finishing time after servicing node v , memory level after servicing v , reduced cost for the path from α to v (sum over all the node weights along the path, for an observation node $[a_{ikq},b_{ikq}]$, the weight $w=w_i-\lambda_i$, for a download node $[a_p,b_p], p\in W$, $w=-\tau_p$), unreachable node subset considering T and time compatibility between v and other nodes, and visited target subset along the path. The sub-problem is therefore to find a longest path from α to β . Since our resolving method is labeling-based dynamic programming, we define a dominate relationship $\boldsymbol{R}$ between two label L_1, L_2 for node v . If $T_1\le T_2, M_1\le M_2, RC_1\ge RC_2$ and there is at least one strict inequation of the three, then $L_1\boldsymbol{R}L_2$. If $L_1\boldsymbol{R}L_2$, then we can

neglect L_2, because it is impossible to generate a better path from v to β based on L_2 than based on L_1. The dynamic programming procedure is as follows:

Step0: Initialization. Attach a label $L_\alpha = (0,0,0,\phi,\phi)$ to α. Label set $\boldsymbol{L} = \{L_\alpha\}$. Every label has a binary flag f indicating whether it is already processed, and $f(L_\alpha) = false$, representing not processed yet.

Step1: Find a non-dominated label and extend all the labels attached on the same node. Finding a label $L_v = (T,M,RC,U,S)$ in $\boldsymbol{L}$ with $min(T)$ and $f(L_v) = false$, if such a label cannot be found, stop; otherwise extend all the labels on v to $V \setminus U \setminus \{v\}$, denote f of all the labels on v as $true$.

Step2: Generate labels and drop dominated ones. For every $g \in V \setminus U \setminus \{v\}$, generate a new label $(T + s_{vg} + d_g, M + m_g, RC + w_g, U_g, S_g)$ for g into $\boldsymbol{L}$, where s_{vg}, d_g, m_g, w_g are transition time between v and g, processing duration of g, memory occupation of g (if g is an observation, then $m_g = d_g$, otherwise $m_g = -d_g$) and weigh of g respectively. U_g is the updated unreachable node subset for g considering $T + s_{vg} + d_g$ and time compatibility; S_g is the visited target subset along the path from α to g via v. Note that all the observation time windows for any element in S_g is in U_g, preventing any cycle of visited targets. If generated label L_g dominates existing labels of g, then delete dominated labels, $f(L_g) = false$; if L_g is dominated by existing labels of g, then delete L_g. Go to step1.

3.3 Branch-and-Price Framework

The branch-and-price framework for the problem is as follows:

Init: Run the FIFO strategy as in [1] to generate a partial schedule for every satellite, and thus build the initial constraint matrix for SP.

Solving SP: Run CPLEX to solve SP, and get the dual variables required.

Solving LPSTRC: Run labeling-based dynamic programming to solve LPSTRC. if inequation $(15) > 0$, then transform the solution to a new column to SP, go to **Solving SP**; otherwise the linear-relaxation of SP is optimally solved, and if the relaxed solution is fractional, then instead directly branching on SP's binary variables, we branching on z_{ijk} of LPSTRC, then go to **Branching**; otherwise we get the optimal solution to the original problem.

Branching: $z_{ijk} = 0$ implies that the arc (i,j) is removed in G; whereas $z_{ijk} = 1$ implies that $z_{ihk} = 0$ for $h \neq j$ and $z_{hjk} = 0$ for $h \neq i$, and the corresponding arcs are removed in G. The above two branches correspond to two distinct LPSTRC, then go to **Solving LPSTRC** for the two LPSTRC respectively.

4 Computational Results

We test our model and algorithm on a 2.4 GHz Intel Core(TM)2 Duo CPU computer. The scheduling scenario information is as follows: 2 satellites (HJ-1-A and HJ-1-B) or 3 satellites (plus HJ-1-C, these 3 satellites form the First Chinese Environment Monitoring Constellation), target number is picked in $[20, 80]$, targets are randomly generated within the latitude $[20, 50]$ and the longitude $[70, 130]$, target rewards are randomly generated in $[1, 10]$, scheduling period is from 1, May, 2010, 12:00 to 3, May, 2010, 12:00. Download windows are generated between the 3 satellites and 2 ground stations in China. Observation span for all the targets and transition time between any two consecutive activities are both set to 30 seconds.

Table 1 Computational results for test scenarios

Instance Name	CPU Time (s)	Solution Value	Is Optimal
2sat_20target	0.95	92	Y
2sat_30target	8.14	150	Y
2sat_50target	29.93	168	Y
3sat_50target	19.82	171	Y
3sat_60target	12.95	141	Y
3sat_80target	600.00	182	N

References

1. N. Bianchessi and G. Righini. Planning and scheduling algorithms for the COSMO-SkyMed constellation. *Aerospace Science and Technology*, 12(7): 535–544, 2008.
2. S. D. Florio. Performances optimization of remote sensing satellite constellations: a heuristic method. In *Proc. of the 5th International Workshop on Planning and Scheduling for Space*, 2006.
3. J. Frank, A. Jonsson, R. Morris, and D. Smith. Planning and scheduling for fleets of earth observing satellites. In *Proc. of the 6th International Symposium on Artificial Intelligence, Robotics and Automation for Space*, 2001.

The Joint Load Balancing and Parallel Machine Scheduling Problem

Yassine Ouazene, Faicel Hnaien, Farouk Yalaoui, and Lionel Amodeo

Abstract The addressed problem in this paper considers the joint load balancing and parallel machines scheduling problem. Two decisions are taken at once: to build the best schedule of n jobs on m identical parallel machines in order to minimize the total tardiness and to find the equitable distribution of the machine's time activity. To our knowledge, these two criteria have never been simultaneously studied for the case of parallel machines. The considered problem is NP-hard since the problem with only the total tardiness minimization is NP-hard. We propose an exact and an approached resolution. The first method is based on the mixed integer linear programming method solved by Cplex solver. The second one is an adapted genetic algorithm. The test examples were generated using the schema proposed by Koulamas [3]for the problem of total tardiness minimization. The obtained results are promising.

Keywords— parallel machine scheduling, total tardiness, load balancing, mixed integer linear programming, genetic algorithm.

1 Introduction

This paper deals with the identical parallel machine scheduling. The objective functions are minimizing total tardiness and balancing workload among the machines. The problem of total tardiness on parallel machine with load balancing is NP-hard, since the problem ($m//\sum T_i$) has been proved to be NP-hard [3]. Rajakumar et al. [6, 7] studied the load balancing problem on parallel machine context. They proposed a genetic algorithm and three heuristics: Longest Processing Time (LPT), Shortest Processing Time (SPT) and Random.

Y. Ouazene, F. Hnaien, F. Yalaoui and L.Amodeo, ICD-LOSI, University of Technology of Troyes
12 rue Marie Curie, 10010 Troyes, France
e-mail: {yassine.ouazene, faicel.hnaien, farouk.yalaoui, lionel.amodeo}@utt.fr

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_79,

Yıldırım et al. [13] proposed a genetic algorithm to solve the problem of minimizing total completion time with load balancing and sequence dependent setups in a non-identical parallel machine configuration. Yu et al. [14] proved that Lagrangian Relaxation found the best solution for the objective of balancing the workload among unrelated parallel machines in a Printed Wiring Board. Sun et al. [10] proposed a genetic algorithm for optimizing two objectives: minimizing the number of pickups and balancing the workload between two gantries in Printed Circuit Board manufacturing industry. Gung and Steudel [2] considered the two criteria of setup time reduction and load balancing in a group technology flow-line cell. Also, Tiwari and Vidyarthi [11] proposed a genetic algorithm-based heuristics to minimize system imbalance on Flexible Manufacturing System (FMS) problems. Kumar et al. [4] studied the Machine-loading problem of a flexible manufacturing system. They extended the simple genetic algorithm-based heuristics and proposed a new methodology: constraint-based genetic algorithm. Stecke [9] addressed several loading criteria like balancing the processing time, workload per machine and minimizing the part movements from machine to machine. Shanker and Rajamathanandan [8] addressed similar kinds of problems. As an extension of these problems, Modi and Shanker [5] solved the loading problem with the objective of part-movement minimization and balancing the workload among machines. They used the Branch-and-Bound procedure and 0-1 integer programming.
The aim of this paper is to formulate a bi-objective mathematical model to solve the joint load balancing and total tardiness minimization on identical parallel machines. The rest of this paper is organized as follows: in Section 2, we present the notations used and introduce the objective functions in details. Next, the bi-objective mixed integer model is developed. In Section 3, we describe the proposed optimization methods. In Section 4, the computational results are presented for both exact method and genetic algorithm. Finally, Section 5 concludes the paper with some perspectives for further research.

2 Problem Formulation

The problem considered in this paper can be formally described as follows: a set of N independent jobs $\{J_1, J_2, ..., J_N\}$ are scheduled on M identical parallel machines. We assume that each job J_j has a deterministic and integer processing time p_j and a known integer due date d_j . We also assume that all jobs are available at time zero ($r_j = 0, j = 1...N$) and no job preemption is allowed. A bi-objective mixed integer programming model is proposed. The first objective is the minimization of total tardiness. The second one is the machine load-balancing.

Notations:

N	total number of jobs
M	total number of machines
$i, j \in \{0, 1, ..., N\}$	job index where job 0 is a fictitious one which is always sequenced at the first position on a machine
m	machine index
p_j	processing time of job j
d_j	due date of job j
c_j	completion time of job j
$T_j = max\{0, c_j - d_j\}$	tardiness of the job j
S_m	set of jobs scheduled on the machine m
$C^m = \sum_{j \in S_m} p_j$	completion time of machine m

Decision variables:

$$x_{ijm} = \begin{cases} 1, \text{ if job } j \text{ immediately follows job } i \text{ in a sequence on machine } m \\ 0, \text{ otherwise} \end{cases}$$

$$y_{jm} = \begin{cases} 1, \text{ if job } j \text{ assigned to machine } m \\ 0, \text{ otherwise} \end{cases}$$

$$C^{max} = max_{1 \leq m \leq M}\{C^m\}$$

$$C^{min} = min_{1 \leq m \leq M}\{C^m\}$$

Mathematical formulation:

Taking into account that C^m can be rewritten as follows:

$$C^m = \sum_{j \in S_m} p_j \Rightarrow C^m = \sum_{j=1}^{N} (p_j \times y_{jm}), m = 1...M$$

We can formulate the workload imbalance as follows:

$$\begin{cases} (C^{max} - C^{min}) \\ C^{max} \geq \sum_{j=1}^{N} (p_j \times y_{jm}), m = 1...M \\ C^{min} \leq \sum_{j=1}^{N} (p_j \times y_{jm}), m = 1...M \end{cases}$$

Finally, the proposed mixed integer linear programming model is the following:

$$Min\ Z_1 = \sum_{j=1}^{N} T_j \tag{1}$$

$$Min\ Z_2 = C^{max} - C^{min} \tag{2}$$

$$\sum_{j=1}^{N} x_{0jm} \leq 1, m = 1...M \tag{3}$$

$$\sum_{i=0, i\neq j}^{N} \sum_{m=1}^{M} x_{ijm} = 1, j = 1...N \tag{4}$$

$$\sum_{j=1, j\neq i}^{N} x_{ijm} \leq y_{jm}, i = 1...N, m = 1...M \tag{5}$$

$$\sum_{i=0, i\neq j}^{N} x_{ijm} = y_{jm}, j = 1...N, m = 1...M \tag{6}$$

$$\sum_{m=1}^{M} y_{jm} = 1, j = 1...N \tag{7}$$

$$c_j + M_1(1 - x_{ijm}) \geq c_i + p_j, i = 1...N, j = 1...N, i \neq j, m = 1...M \tag{8}$$

$$T_j \geq c_j - d_j, j = 1...N \tag{9}$$

$$c_j > 0, j = 1...N \tag{10}$$

$$T_j \geq 0, j = 1...N \tag{11}$$

$$C^{max} \geq \sum_{j=1}^{N} (p_j \times y_{jm}), m = 1...M \tag{12}$$

$$C^{min} \leq \sum_{j=1}^{N} (p_j \times y_{jm}), m = 1...M \tag{13}$$

$$x_{ijm}, y_{jm} \in \{0,1\}; C^{max}, C^{min} \geq 0; c_0 = 0 \tag{14}$$

In the above model, Equations (1) and (2) are the objective functions, respectively minimizing the total tardiness and minimizing the workload imbalance among the machines. Equation (3) assures that for each machine, only one real job follows the fictitious job 0. Equation (4) states that a job must be processed at one and only one position on a machine and it will be immediately preceded by exactly one job.Equation (5) states that if job i is processed on machine m, it will be immediately followed by at most one other job on this machine. Equation (6) states that job j should immediately follow another job on machine m if it is placed on this machine. Equation (7) guaranties that each job is assigned to exactly one machine. In Equation (8), M_1 is a large positive number. This equation expresses the jobs completion times related to the jobs processing times and positions. Equation (9) represents the relation between the completion time of each job, its due date and tardiness variable. Equations (10) and (11) ensure that completion times and tardiness are positive variables. Equations (12) and (13) represent workload-balancing constraints. Equation (14) states the properties of the decision variables. It states also that the completion time of the fictitious job is equal to zero.

3 Optimization Methods

In this section we propose an exact and an approached resolution methods.The first one is based on the mixed integer linear programming method. We establish a weighted objective function to include the two objective functions in a single one:

$Min\ Z_3 = W_1 \times \sum_{j=1}^{N} T_j + W_2 \times (C^{max} - C^{min})$. W_1 and W_2 are the weight of the objective functions. For our study, we consider the case where $W_1 = W_2 = 0.5$ (the two objective functions have the same importance). Using the weighted objective function and considering the above model constraints, we obtain a single objective mixed integer programming model, which can be solved using Cplex solver. The second one is a genetic algorithm adapted for our problem based on a general version of the genetic algorithm.

4 Computational Results

This section describes the computational results. The test examples were generated using the schema proposed by Koulamas [3] and Azizoğlu and Kırca [1] for the problem of total tardiness on parallel machines. We generate, for each job j, an integer processing time p_j using a uniform distribution [1,100] and an integer due date d_j in the interval $[P \times (1 - TF - (RDD/2)), P \times (1 - TF + (RDD/2))]$, where $P = (\sum_{j=1}^{N} p_j)/M$; $TF, RDD \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Considering all the possible combinations of the two parameters TF and RDD we obtain 25 representative groups of data [12]. Ten size of problems were tested with M = 3 and 4, N = 8, 10, 25, 50 and 75. We have also fixed at 30 minutes the computational time of the Cplex solver. Both Cplex solver and the genetic algorithm give the exact solution of all the instances of the problems with M = 3 and 4, N = 8 and 10. Considering the stochastic character of the genetic algorithm, we have tested its stability among more important instances: (N = 25, M = 4), (N = 50, M = 4) and (N = 75, M = 4). For each instance and for each group of data, 100 independent runs were performed. After that, we calculated the gap between the best and the worst found solutions. The obtained results are interesting, with average gap of, respectively, 1.2%, 1.5% and 2.1%.

5 Conclusion and Further Research

This paper addressed the identical parallel machine scheduling to minimize simultaneously the total tardiness and the workload imbalance among the machines. A bi-objective mathematical formulation is proposed. An exact and an approached methods are proposed. The first one is based on a mixed integer linear programming method solved by Cplex solver using a weighted objective function. The second one is a genetic algorithm.

This study may be extended by determining a lower bound to test the effectiveness of the genetic algorithm among more important instances.
As a part of future research an exact and an approached multi-objective methods to solve the considered problem are under construction. The first one is based on the Two Phases Method and the second one is a multi-objective genetic algorithm.

References

1. M. Azizoğlu and O. Kırca. Tardiness minimization on parallel machines. *International Journal of Production Economics*, 55: 163–168, 1998.
2. R.R. Gung and H.J. Steudel. A workload balancing model for determining set-up time and batch size reductions in GT flow line work cells. *International Journal of Production Research*, 37(4): 769–791, 1999.
3. C. Koulamas. The total tardiness problem: review and extensions. *Operations Research*, 42: 1025–1041, 1994.
4. A. Kumar, Prakash, M. Tiwari, R. Shankar, and A. Baveja. Solving machine loading problems in a flexible manufacturing system using a genetic algorithm based heuristic approach. *European Journal of Operational Research*, 175: 1043–1069, 2006.
5. B.K. Modi and K. Shanker. A formulation and solution methodology for part movement minimization and workload balancing at loading decisions in FMS. *International Journal of Production Economics*, 34: 73–82, 1994.
6. S. Rajakumar, V.P. Arunachalam, and V. Selladurai. Workflow balancing strategies in parallel machine scheduling. *International Jornal of Advanced Manufacturing*, 23: 366–374, 2004.
7. S. Rajakumar, V.P. Arunachalam, and V. Selladurai. Workflow balancing in parallel machines through genetic algorithm. *International Journal of Advanced Manufacturing Technology*, 33: 1212–1221, 2007.
8. K. Shanker and S. Rajamathanandan. Loading problemin FMS: part movement minimization. In *Proc. 3rd ORSA/TIMS conf. on flexible manufacturing systems: Operations Research Models and Applications*, pages 99–104, 1989.
9. K.E. Stecke. Formulation and solution of nonlinear integer production planning problems for flexible manufacturing systems. *Management Science*, 29: 273–289, 1983.
10. D.S. Sun, T.E. Lee, and K.H. Kim. Component allocation and feeder arrangement for a dual-gantry multi-head surface mounting placement tool. *International Journal of Production Economics*, 95: 245–264, 2005.
11. M.K. Tiwari and N.K. Vidyarthi. Solving machine loading problems in a flexible manufacturing system using a genetic algorithm based heuristic approach. *International Journal of Production Research*, 38(14): 3357–3384, 2000.
12. F. Yalaoui and C. Chu. Parallel machine scheduling to minimize total tardiness. *International Journal of Production Economics*, 76: 265–279, 2002.
13. M.B. Yıldırım, E. Duman, K. Krishna, and K. Senniappan. Parallel machine scheduling with load balancing and sequence dependent setups. *International Journal of Operations Research*, 4: 42–49, 2007.
14. L. YU, M. Shih, M. Pfund, M.W. Carlyle, and W.J. Fowler. Scheduling unrelated parallel machines: an application to PWB manufacturing. *IIE Transactions*, 34: 921–931, 2002.

III.1 Enviromental Management

Chair: Prof. Dr. Paola Zuddas (University of Cagliari, Italy)

The challenges of Environmental Management call for a strongly interdisciplinary approach, and Operations Research methodologies and techniques may play a fundamental role. Take for instance Land and natural Resources Management. It is an important component of local, national and also international policy. Operation Research may give a crucial support to decision makers in the identification of the best compromise among people, economics, political constraints and environmental needs, in the shaping of the concrete objectives and in finding the path of action which is the most suitable to reach those objectives.

Recent European directives and national political decisions open new problems in environmental management, calling for new policies in resources pricing (water and energy liberalization) and in resources preserving and restoring strategies (scarce resources repair).

Moreover, the many different sources of uncertainty in data identification make the decision process very risky.

Submissions involving one or more of the aspects of the Decision Process in Land and Resources Management are welcome. Of particular interest are issues such as water, energy, traffic congestion, etc.

Papers which help closing the gap between academic research and real life decision making will be particularly appreciated.

Estimating the Short-Term Impact of the European Emission Trade System on the German Energy Sector

Carola Hammer and Gunther Friedl

Abstract This study focuses on the European Emission Trade System introduced in 2005, since in contrast to other regulatory instruments it remains relatively untested in its economic effects. Because of the inherent uncertainty of the price for emission allowances (EAs), we first model the demand and supply function. For former this study employs contribution margin accounting and production functions of the available power generation technologies as well as combined heat and power installations in Germany. The resulting demand function for EAs is relatively price inelastic at some points due to spreads in primary energy prices and the existing production capacities of different technologies. Consequently, the quantity of EAs has an extensive influence on the price of EAs and so on different indicators with economic importance.

1 Introduction

In order to counteract anthropogenic climate change an international target agreement (-8 % of green house gas (GHG) emissions in the EU and -21 % in Germany in the period 2008-2012 in comparison to the level in 1990) was ratified with the Kyoto Protocol in 1997 and came into force 2005. To achieve the targets a package of international, European and national environmental policy instruments was implemented.

Broken down into different gases and sectors in national allocation plans, a reduction of -21 % GHG implies an annual restriction (CAP) of 453 m. t CO_2 (that means 453 m. emission allowances (EA) in the European Emission Trade System (ETS)) for the energy sector and energy intensive industry during the period of 2008-2012. Since these two sectors have to buy an EA for every emitted ton of CO_2, thus

Carola Hammer and Gunther Friedl
Chair of Management Accounting and Control, Technical University Munich, Arcisstrasse 21, 80333 Munich, e-mail: Carola.Hammer@wi.tum.de or Gunther.Friedl@wi.tum.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_80,

the ETS internalizes environmental costs by adding a new cost type to the objective function of the companies.[1]

The economic literature on design and institutional issues of the ETS is wide but econometric research is lacking due to the fact that the introduction of the ETS only happened in 2005 [2]. Here, we show the response of the enterprises on the new cost factor of the ETS by using a production planning model as well as technical and market data from the German energy sector. Because of the existence of combined heat and power generation plants (CHP), both the electricity market, and the heat market are taken into account. In contrast to other studies we also respect the difference between EAs and taxes in taking the EA price as uncertain variable.

Since the ETS is only one instrument from a package of different environmental policy instruments, we also consider the feed-in tariffs (price guarantee) of the Erneuerbare Energien Gesetz (EEG) for renewables and the surcharge on the market price of the Kraft-Waerme-Kopplungs-Gesetz (KWKG) for CHPs (subsidies). This requires a maximization of profit instead of a minimization of costs and use of contribution margin accounting in place of a marginal cost curve, as in many other production planning models.

After setting up the model and applying a solution method we analyze, whether the ETS fulfills the requirements of policy makers and the OECD criteria "ecological effectiveness and economic efficiency".[2] In addition we consider the impact of the ETS on other national objectives, as independence of energy imports, low consumer prices for power and heat and power generation in energy efficient CHP, as well as the burden of the ETS on the energy sector.

2 Model Formulation

2.1 Objective Function

The short-term objective function of production planning in the German energy sector is to maximize the annual profits, which consist of the profits from electricity and heat generation as well as profits from power exchange within the Union for the Coordination of the Transmission of Electricity (UCTE).

$$\max G = \sum_{i=1}^{I} G_i + G_h + G_n \tag{1}$$

The profit per power plant or per combined heat and power (CHP) plant (i) is the contribution margin referring to the amount of power multiplied by the amount of generated power ($P_{el,i} \cdot t_i$) in the existing capacities ($P_{el,i}$) minus fixed costs

[1] This fact distinguishes European studies on energy systems from those from other countries, where emission abatement is not an issue or where there is a legal requirement on emissions [6].

[2] System security [5], distribution [1], capacity expansion and location decisions [6], as well as EAs from a trading perspective [4] are not within the scope of this study. Where necessary, we refer to the literature in the Appendix.

($K_{F,i}$). The contribution margin is calculated by subtracting the variable costs of fuel ($k_P/\eta_{el,i}$), emissions ($k_E \cdot EF_{el}$), and operation and maintenance ($k_{OM,i}$) from the revenue. Here, main cost drivers are the prices for primary energy (k_P) and EAs (k_E).

The revenue includes subsidies for renewable energy ($p_{EEG,i}$) and CHP ($p_{KWKG,i}$) as well as the sales of power (p_{el}) and heat (p_{th}). The amount of generated heat is based on the power production and calculated by the power ratio of a plant (σ_i)[3].

$$G_i = [\max(p_{EEG,i}, p_{el} + p_{KWKG,i}) + p_{th} \cdot (1-\sigma_i)/\sigma_i - k_P/\eta_{el,i} - k_E \cdot EF_{el} - k_{OM,i}] \cdot P_{el,i} \cdot t_i - K_{F,i} \quad (2)$$

Since the demand constraint for heat (D_{th}) has to be fulfilled, heat, which is not produced by CHP, has to be generated by heat only plants.[4] Here sales, variable costs, efficiency factor ($\eta_{th,h}$) and emission factor (EF_{th}) refer to the amount of heat.

$$G_h = (p_{th} - p_P/\eta_{th,h} - EF_{th} \cdot k_E - k_{OM,h}) \cdot (D_{th} - \sum_{i=1}^{I} (1-\sigma_i)/\sigma_i \cdot P_{el,i} \cdot t_i) - K_{F,h} \quad (3)$$

In contrast to heat distribution, the losses in power transmission are lower, which offers the opportunity to import or export power. This leads to a flexibility in producing more or less power than the domestic demand for power (D_{el}). Whether it is advantageous to import or to export depends on the price difference of the domestic (p_{el}) and the foreign electricity price ($p_{el,foreign}$) and so on the difference in the technology portfolios.

$$G_{ForeignTrade} = (p_{el} - p_{el,foreign}) \cdot (D_{el} - \sum_{i=1}^{I} P_{el,i} \cdot t_i)) \quad (4)$$

2.2 Constraints

The produced power plus the balance of import and export (net import) must meet the domestic power demand. The range of importable or exportable power amount depends on the available grid capacity at the borders ($P_{el,b}$).

$$|D_{el} - \sum_{i=1}^{I} P_{el,i} \cdot t_i| \leq 8,760 \cdot \sum_{b=1}^{B} v_b \cdot P_{el,b} \quad (5)$$

The sum of generated heat in CHP as well as in heat only plant production capacities has to fulfill the demand function for heat, but must not require a heat production

[3] The share of electricity to total (electric plus thermal) capacity lies between zero and one. According to the definition, the power ratio is zero for a heat only plant and one for a power only plant. CHP lies in between zero and one ($0 < \sigma_i \leq 1 \quad \forall i$).

[4] Since the heat only plants (h) are in the minority, and not the main focus of our analysis, their capacities ($P_{th,h}$) are modeled as one big plant.

above the available capacities.

$$0 \leq D_{th} - \sum_{i=1}^{I} (1-\sigma_i)/\sigma_i \cdot P_{el,i} \cdot t_i \leq v_h \cdot P_{th,h} \cdot 8,760 \tag{6}$$

In one year the maximal running time (t_i) of a plant is 8,760 h multiplied by its availability (v_i) [5]. The availability of a plant depends on down time, revision time and, in the case of renewable energy, on the availability of the primary energy.

Switch off of base load plants involves revisions, and, in addition, start-up and revision costs for these plants are quite high. Therefore they will only be shut down completely in exceptional cases. Hence base load plants run at least at a certain minimum of full load hours ($\underline{t}_i$) in our model. To avoid underestimation the same holds for peak load plants.

$$\underline{t}_i \leq t_i \leq 8,760 \cdot v_i \quad \forall i \tag{7}$$

3 Solution Method and Results

With the introduction of ETS, a new cost factor is included in the contribution margin accounting of the energy sector. Therefore the energy sector has to rearrange its merit order correspondingly. By maximizing its profits, the energy sector additionally receives its optimal quantity of carbon emissions. Due to the linearity of the problem, we program the production model in ILOG OPL and solve it with a simplex algorithm.[6]

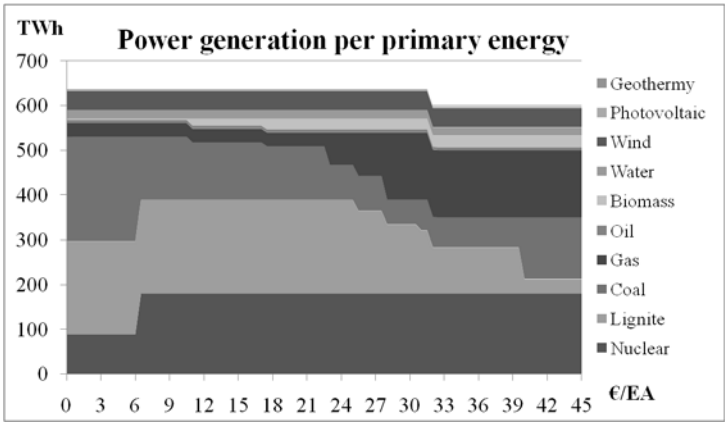


Fig. 1 Power generation per primary energy as a function of the EA price

[5] The availability calculated as quotient of available annual full load hours of a plant to 8,760 h is located between zero and one ($0 \leq v_i \leq 1 \quad \forall i \quad and \quad 0 \leq v_h \leq 1$).

[6] In the first step, the simplex algorithm is used to find a feasible solution for the full load hours of each plant and in the second step for an optimal feasible solution.

In contrast to taxes and subsidies, the price for EAs is not exogenously given, but a result of supply and demand. Therefore we have to calculate the demand function for EA with a sensitivity analysis first. Since the switch load effect (Fig. 1) leads to a discontinuous slope in emission quantities and so in the demand function, our model is used in an iteration method while raising the price of EAs.

By adding the annual CAP of the ETS to the resulting demand functions (Fig. 2), the price for one EA is arrived at. Since the option for banking and borrowing the EAs within a trading period exists, we have to determine all years of a trading period and calculate the average to smooth price effects of demand shifts as observed in the years 2005-2007.

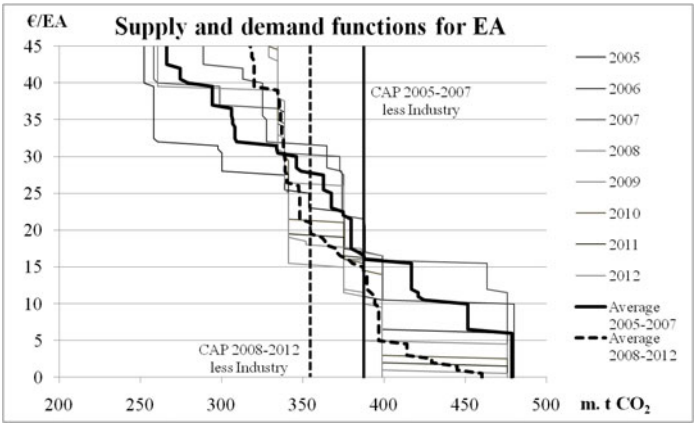


Fig. 2 Supply and demand functions of the energy sector for EA in 1. and 2. trading period

After the estimation of the probable price range for EAs, we are able to compute some indicators to assess the impacts of the ETS. The emission quantity and emission costs per output unit should allow a first evaluation according to the OECD criteria. In the first trading period, the ETS leads to an annual emission reduction of 79 m. t CO_2 and thus an emission amount per output unit of 0.628 t CO_2 /MWh$_{el}$ instead of 0.753 t CO_2 /MWh$_{el}$ without ETS. In the second period, the annual emission reduction of 107 m. t CO_2 implies 0.568 t CO_2 /MWh$_{el}$ in place of 0.751 t CO_2 /MWh$_{el}$.

These emission reductions cost the energy sector 6,349.5 m. Euro/a in 2005-2007 and 6,873.4 m. Euro/a in 2008-2012. This corresponds to the internalized part of the environmental costs. The ratio of emission costs to sales or to contribution margin gives an estimate for the burden of the ETS on the energy sector. The former rises from 0 without ETS to 13.19 % with ETS in 2005-2007 and to 11.89 % in 2008-2012 and the latter increases from 0 to 26.87 % in the first trading period and to 23.18 % in the second trading period.

In a competitive market the emission costs per output unit are comprehensively reflected in the power price [2]. Seeking to guarantee economical power prices for households and industries, national policy makers are interested in low emission

costs per power output unit. In the first trading period, the emission costs per power output unit amount to 9.98 Euro/MWh $_{el}$ and their share in power price to 19.96 %. In the second trading period they are 11.04 Euro/MWh $_{el}$ and 17.64 %. Since emission costs per output unit are not low, this explains the significant rise of power prices with the introduction of the ETS in the year 2005 [3]. Furthermore, due to the rise of energy prices the share of emission costs in power price, in sales or in contribution margin decreases in the second period in comparison to 2005-2007, although the average EA price and the annual emission costs increase.

Finally, we demonstrate an influence of the EA price on the dependence on energy imports and power generation in CHP by computing the output per primary energy or technology. Since domestic lignite has a high emission factor, the ETS implies an increase in import dependency. However, this effect barely kicks in until an allowance price of 28 Euro due to the price spread between the different fuels. The same effect can be observed for the share of power generation in CHP, which noticeably increases from a price of 28 Euro.

4 Discussion and Outlook on Future Research

The demand function shows a variation in the price elasticity for EAs. On some points on the demand function, a small decrease of the EA quantity heavily increases the price for EAs and so the costs for the energy sector and power consumers, while on other points a large tolerance exists. This effect requires a careful selection of the EA amounts for the policy makers, so that power prices will not be uselessly increased and "ecological effectiveness and economic efficiency" are respected.

Besides the quantity of EAs, the development of the energy prices and capacities of different technologies have an extensive influence on the price of EAs. Therefore future research should focus on the long-term impacts, such as the question of how the ETS can affect investment decisions.

References

1. W. Fichtner. Emissionsrechte, Energie und Produktion Verknappung der Umweltnutzung und produktionswirtschaftliche Planung. Habilitation der Universität Karlsruhe, Berlin, 2005.
2. F. Gulli. *Modelling the short-run impact of carbon trading on the electricity sector.* Edward Elgar Publishing Limited, Cheltenham, Northampton, 2008.
3. A. Ockenfels. Strombörse und Marktmacht. *Energiewirtschaftliche Tagesfragen*, 57(5): 44–58, 2007.
4. A. Rong and R. Lahdelma. CO2 emissions trading planning in combined heat and power production via multi-period stochastic optimization. *European Journal of Operational Research*, 176: 1874–1895, 2007.
5. H. Roth. *Modellentwicklung zur Kraftwerksparkoptimierung mit Hilfe von Evolutionsstrategien.* Dissertation der Technische Universität München, München, 2008.
6. J. Sirikum, A. Techanitisawad, and V. Kachitvichyanukul. A New Efficient GA-Benders Decomposition Method: For Power Generation Expansion Planning with Emission Controls. *IEEE Transactions on Power Systems*, 22(3): 1092–1100, August 2007.

The Water Pricing Problem in a Complex Water Resources System: A Cooperative Game Theory Approach

G. M. Sechi, R. Zucca, and P. Zuddas

Abstract The research presents a methodology to allocate water services costs in a water resources system among water users using a Cooperative Game Theory approach based on the integral river basin modelling. The proposed approach starts from the characterization of the system to be modelled. The Decision Support System WARGI [7, 4, 6] is then used to achieve the best water system performances and to calculate the least cost of each one of the users' coalitions that may arise using the resources. The cost allocation is evaluated by the Cooperative Game Theory methods. The aim of the work is to suggest a tool for decision makers to define water price policies in accordance with the sustainability and fairness principles settled by European Water Framework Directive [2].

1 Introduction

A central problem in planning the provision of public services when dealing with limited resources is how to determine a "fair" and "just" allocation of management costs. This problem is particularly relevant for water systems in Europe to comply with the Water Framework Directive 2000/60/CE [2], which addresses the recovery costs of water services considering adequate contributions and priorities when dealing with different water uses. The actual water pricing methods are mainly based on countable or historical cost (sunk costs) allocation corresponding to previous investments, and they are used as a simple cost recovery instrument.

G. M. Sechi
University of Cagliari, Via Marengo 2, 09123, Cagliari, Italy, e-mail: sechi@unica.it

R. Zucca
University of Cagliari, Via Marengo 2, 09123, Cagliari, Italy, e-mail: rzucca@unica.it

P. Zuddas
University of Cagliari, Via Marengo 2, 09123, Cagliari, Italy, e-mail: zuddas@unica.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_81,

The cost allocation criterion is generally determined by legal imposition, and the users do not make any decisions, or they are simply consulted about the possible alternatives. Otherwise, a Cooperative Game Theory (CGT) approach is able to determine a sustainable, acceptable, rational and fair cost-sharing rule [12] and it is particularly appropriate for contexts like water services, in which it is important to define the agreements and to encourage cooperation among decision makers.

2 Cooperative Game Theory

CGT belongs to the mathematical science called Games Theory, developed in the last century by [10], and it searches cooperative solutions studying the individual decisions in situations in which there are some interactions among different decisional subjects. There are several applications of CGT also concerning water resources [9, 13, 3, 1]. To define a cooperative game we have to explain the following definitions, as in [12]. Let $N = \{1,2,3,\ldots,n\}$ be a set of players, for example different water users in a water resources system. Every subset S of N is called "coalition" and N is the "grand coalition". Let $c(i)$ be the cost of providing player i by itself and $c(S)$ the cost of providing players in S jointly. Consequently, $c(N)$ is the cost of grand coalition, i.e. in our contest the cost related to whole water service. By convention $c(\Phi) = 0$. The cost associated with a coalition represents the least cost to supplying the player in that coalition and the discrete function constituted by the least costs of every coalition is the so-called "characteristic function". An allocation is a vector $(x_1, x_2, \ldots, x_n)$ such that

$$\sum_N x_i = c(N) \,, \tag{1}$$

where x_i is the amount charged to player i. Allocations should respect two fundamental principles: the rationality and the marginality principle. By the first one, if cooperation among players is voluntary, then self-interest dictates that no participant Ð or group of participant Ð be charged more than their stand alone (opportunity) cost:

$$\sum_S x_i \leq c(S) \,, \tag{2}$$

otherwise they would have no incentive to agree to the proposed allocation. The second principle states that no player should be charged less than his marginal cost of including him in a coalition:

$$\sum_S x_i \geq c(N) - c(N-S) \,, \tag{3}$$

otherwise it could be said that the coalition $N-S$ is subsidizing S. The condition (2) provides incentives for voluntary cooperation and the condition (3) arises from considerations of equity. Given the formula (1) the conditions (2) e (3) are equiva-

lent. There are two different types of solution of a cooperative game. The first one is the set of admissible solutions, the so-called "core". The core of a game is the set of all allocations $x \in \mathbf{R}^{\mathbf{N}}$ such that (1) and (2), or equivalently (3), hold for all S of N [11]. The second type is represented by a single allocation. One of the most utilized single allocation method is the Shapley Value that responds to symmetry, additivity and monotonicity principles [8]. The Shapley Value is defined by this formula:

$$x_i = \sum_{S \subseteq N-i} \frac{|S|!\,(|N-S|-1)}{|N|!}\,[c(S+i)-c(S)]\ , \tag{4}$$

where $|S|$ is the cardinality of coalition S, i.e. the number of players involved in the coalition and, consequently, $|N| = n$.

3 Cost Allocation Methodology

The cost allocation methodology applied to water systems consists of the following main steps:

1. Water resource system analysis: analysis of hydrological, hydraulic, economic and environmental aspects of the system;
2. Cooperative Game definition: identification of independent agents and definition of coalitions;
3. Characteristic function calculation: evaluation of the least cost for every coalition using the optimization model WARGI;
4. Game solution: application of the CGT allocation methods.

The evaluation of the characteristic function is the base of cooperative games and requires a cost analysis associated with each possible coalition system, which implies an optimisation process whose magnitude grows exponentially with the number of system agents. The proposed approach starts characterizing the hydrologic, hydraulic, economic and environmental aspects of the system to be modelled. The characterization of the water system includes extended time-horizon data of surface hydrology, given by monthly runoff time series, the application of continuity equations and balance equations in the reservoirs and aquifers nodes. The determination of water costs is based on the construction, management and operation cost functions for the hydraulic infrastructures. The Decision Support System (DSS) WARGI [7, 4, 6] to optimize water resources systems is then used to achieve the best water system performances and to calculate the least cost of each one of the users' coalitions that may arise using the resources. The outputs of the optimization process give the characteristic function of the game and, so, it is possible to apply a CGT method to evaluate the cost-allocation.

3.1 CGT Application

The methodology is been applied in Flumendosa - Campidano water system in Sardinia, Italy. The system (Fig. 1) confers the resource to the three main users: municipal, irrigation and industrial, that can be supply using different infrastructures: dams, diversion site, channel and pipelines. Moreover, an interconnection to another water system with a pumping station exists. The water demands are reported in Fig. 2. In Flumendosa - Campidano application we considered the OMR (operating, maintenance and replacement) costs of the infrastructures given by the regional document Piano Stralcio di Bacino [5], given in Table 1. By the DSS WARGI the characteristic function is been estimated, as we report in Table 2. Through the characteristic function is possible to estimate the core of the game, calculating the minimum and maximum costs of each user (Table 3).

Consequently the core of the game is defined by the following mathematical system:

$$\begin{cases} Mun. + Irr. + Ind. = 533.48 \\ 0.00 \leq Mun. \leq 469.08 \\ 41.66 \leq Irr. \leq 533.48 \\ 1.42 \leq Ind. \leq 185.11 \end{cases} \tag{5}$$

Moreover, we are able to give a single allocation reference using the Shapley Value (Table 3).

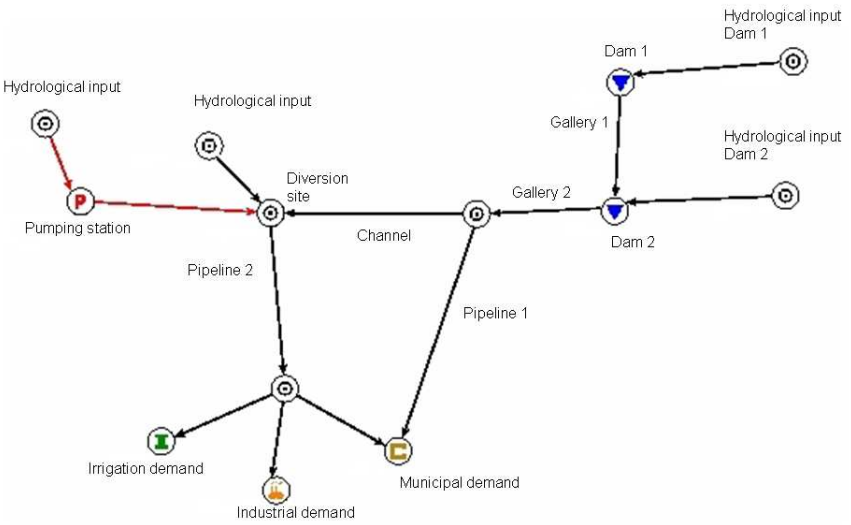


Fig. 1 Graph representing the water resource system.

Table 1 OMR costs of the infrastructures

Infrastructures	**OMR costs [€/ year]**
Dam 1	523748
Dam 2	498878
Gallery 1	64331
Gallery 2	116604
Channel	353951
Pipeline 1	874116
Diversion site	14000
Pipeline 2	1926526
Interconnection + Pumping station	873979

Table 2 Characteristic function

Coalitions	Mun.	Irr.	Ind.	Mun. + Irr.	Mun. + Ind.	Irr. + Ind.	Grand coalition
Cost [M€]	469.08	533.48	185.11	532.06	491.82	533.48	533.48

Table 3 Maximum and minimum values for each user and Shapley value (in M€)

Users	Minimum value	Maximum value	Shapley Value
Municipal	0.00	469.08	207.24
Irrigation	41.66	533.48	260.27
Industrial	1.42	185.11	65.97

4 Conclusions

The methodology, based on CGT, could be a valuable tool able to define water price policies and economical analyses that the water districts have to realize according

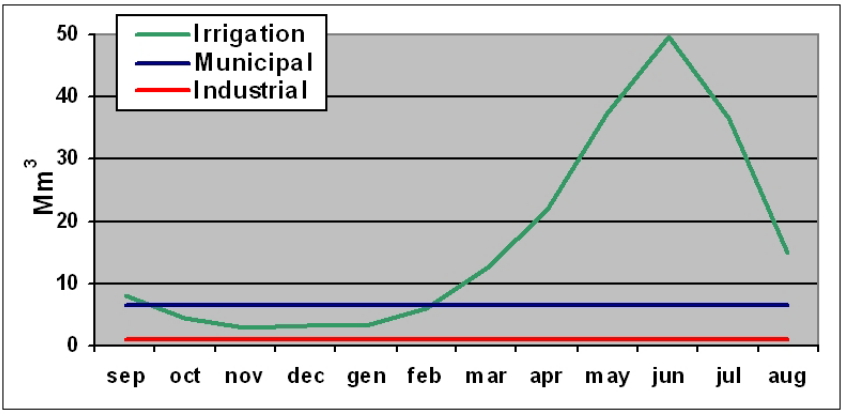


Fig. 2 Water demands.

to the European Water Framework Directive. The evaluation of the characteristic function is the base of cooperative games and it requires an optimization process through the use of DSS WARGI.

References

1. D. Deidda, J. Andreu, M. A. Perez, G. M. Sechi, R. Zucca, and P. Zuddas. A cooperative game theory approach to water pricing in a complex water resource system. In *Proceedings of the 18th World IMACS/MODSIM Congress*, Cairns, Australia, 2009.
2. European Union. Directive 2000/60/EC of European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. 2000.
3. I. Lippai and J. P. Heaney. Efficient and equitable impact fees for urban water system. *Journal of Water Resources Planning and Management*, March 2000.
4. A. Manca, G. M. Sechi, and P. Zuddas. Complex water resources system optimization aided by graphical interface. In *Proceedings of the VI International Conference of Hydroinformatics*, Singapore, 2004.
5. Regione Autonoma della Sardegna. Piano stralcio di bacino regionale per l'utilizzo delle risorse idriche, 2006.
6. G. M. Sechi and A. Sulis. Water System Management Through a Mixed Optimization-Simulation Approach. *Journal of Water Resources Planning and Management*, 3(135): 160–170, 2009.
7. G. M. Sechi and P. Zuddas. WARGI: Water resources system optimization aided by graphical interface. In *Hydraulic Engineering Software*, pages 109–120. WIT-PRESS, 2000.
8. L. S. Shapley. A value for n-person games, Contributions to the Theory of Games. In *Annals of Mathematics Studies*. Princeton University Press, 1953.
9. Tennessee Valley Authority. Allocation of investment in Norris, Wheeler and Wilson Project, 1938.
10. J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton Uiversity Press, Princeton, New Jersey, USA, 1944.
11. H. P. Young. *Cost allocation: methods, principles, applications*. Elsevier Science Publishers, Princeton, New Jersey, USA, 1985.
12. H. P. Young. *Handbook of game theory with economic application*. Elsevier Science Publishers, 1994.
13. H. P. Young, N. Okada, and T. Hashimoto. Cost allocation in water resources development. *Water Resources Research*, 1(18): 463–475, 1982.

III.2 Energy Markets

Chair: Dr. Hans-Georg Zimmermann (Siemens, München)

Today's primary energy markets are high interrelated and can be understood as large interacting dynamical systems. As all markets they depend on the supply and demand side.

In the past the demand side was the uncertain part in this dynamics and one task in the optimization of the system was the forecast of the demand. With the upcoming renewable energy not only the demand but also the supply side has to be forecasted and the interaction has to be optimized.

Characteristics of energy markets are high dimensionality (because of the interaction of the different energy markets), nonlinearity (because of the decision making of traders) and different time scales of trading behavior and long term makroeconomical effects. In this section we want to study techniques and models, that help to optimize energy markets.

Analysis of Electrical Load Balancing by Simulation and Neural Network Forecast

Cornelius Köpp, Hans-Jörg von Mettenheim, Marc Klages, and Michael H. Breitner

Abstract The rising share of renewable energy poses new challenges to actors of electricity markets: wind and solar energy are not available without variation and interruption, so there is a rising need of high priced control energy. Smart grids are introduced to deal with this problem, by load balancing at the electricity consumers and producers. We analyze the capabilities of electrical load balancing and present initial results, starting with a short review of relevant literature. Second part is an analysis of load balancing potentials at consumer households. A software prototype is developed for simulating the reaction to dynamically changing electricity rates, by implementing two generic classes of smart devices: devices running once in a defined limited time slice, as as simplified model of real devices like dish washer, clothes washer, or laundry dryer; devices without time restriction and a given daily runtime, as a simplified model of water heater with large storage. Third part is an analysis of centrally controlled combined heat and power plants (CHPP) for load balancing in a virtual power plant composed of CHPPs, wind and solar energy plants. CHPP load is driven by heating requirements but we want to forecast the (uninfluenced) produced electricity. Our neural network forecast of CHPP load allows to alter the behavior of heat (and electricity) production. In times of low demand or high production by wind and solar energy, the CHPP can be switched off, provided that sufficient heat reserves have been accumulated before. Based on the neural network forecast, a software prototype simulates the effects of load balancing in virtual power plants by controlling the CHPPs.

Cornelius Köpp, e-mail: koepp@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

Hans-Jörg von Mettenheim, e-mail: mettenheim@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

Marc Klages, e-mail: klages@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

Michael H. Breitner, e-mail: breitner@iwi.uni-hannover.de
Institut für Wirtschaftsinformatik, Leibniz Universität Hannover

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_82,

1 Introduction

The price of a commodity on the free market is determined by the relationship between supply and demand. The demand for energy (understood as energy consumption), on the contrary, fluctuates permanently on the market. During the day more power is needed than at night. Especially around lunch time and in winter evenings the demand is high. Although energy supply is regulated according to fluctuating demand, the proportion of uncontrollable variation on the supply side increases steadily. This is due to the development of renewable sources of fluctuating energy such as wind turbines and photovoltaic systems. Currently, electricity rates for end users are fix. Thus, this cannot help to compensate for the occurring fluctuations of supply and demand. To circumvent the problem on the long run, variable electricity rates can provide the customers with an incentive to shift their electricity consumption to times in which, for example, a surplus of wind energy is given. On the short run, potentials are raised through the promotion of decentralized capacity to produce electricity, heat and cold, below the 20 MW threshold (CHPP). In addition, producers can diversify supply according to demand as well as weather and climate data. The load shifts on the part of the energy producers can help to relieve the nets and smoothen the supply and demand over time.

Considering the increasing number of journals, workshops, conferences and even books, the systematic literature review has become an indispensable method for investigation of special topics [5]. Following up on [4], we use an analytical framework for the analysis of internationally published articles which affect the issue of forecast optimization techniques for electrical load balancing. A systematic literature review is limited by the paper selection process and the quality of the input material. However, in the underlying approach, the risk has been minimized by following a proven method for creation of a literature review by [11].

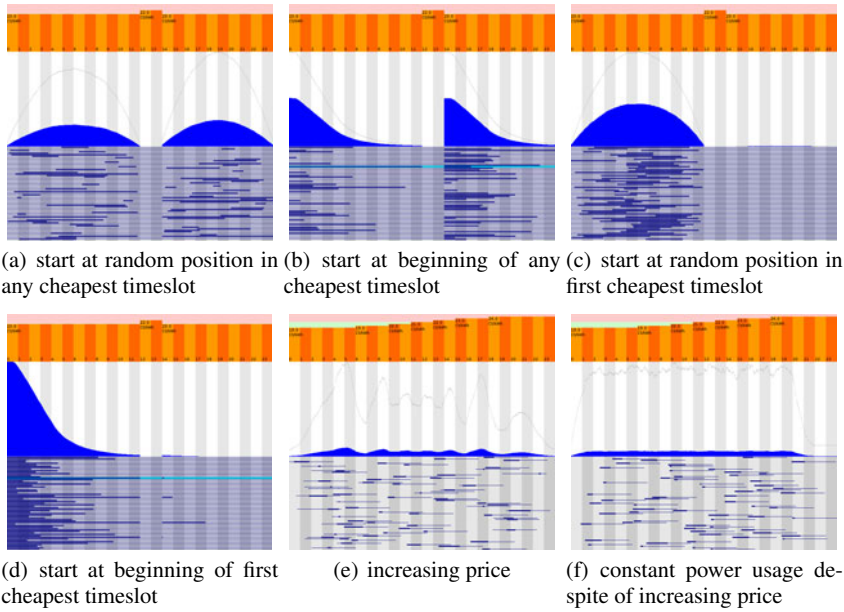
In total we have analyzed over 90 sources in detail, especially by method, content and significance. Due to the limited space in this article the results of the literature analysis is presented briefly. The literature analysis showed that there is not much publicly accessible literature which handles the term "load balancing/shifting/control and prognosis methods" [6, 8]. In addition, to the economic character [7] of the energy market (like price elasticities of consumers [9] and game theoretical analysis of market behaviors of individual participants [1]), the literature more often discusses the essential structural changes needed to establish a decentralized energy network [12], combining smart metering [3, p.145], virtual power plants [13] and a smart market like the EEX today [10]. Furthermore, many authors analyze the necessary IT changes to establish this vision [2, p.382].

The lack of methods to predict a necessary load balancing is clear and justifies the research presented here.

2 Load Balancing Potentials at Consumer Households

The power usage of consumer households shall be controlled by a dynamic electricity rate. This concept assumes that low/high rates will increase/decrease power usage to allow a shifting of power consumption. The rate implemented by the simulation prototype is fixed at 0:00 for 24 hours in advance, and the price (between a defined minimum and maximum) is constant for at least 2 hours.

We identified two classes of smart devices: 1) Devices running once (for few minutes to several hours) in a defined limited time slice (of several hours, typically much longer than running time), as simplified model of real devices like dish washer, clothes washer, or laundry dryer. 2) Devices without time restriction and a given daily runtime, as a simplified model of water heater with large storage. The daily running time is based on running time of real water heaters.



(a) start at random position in any cheapest timeslot

(b) start at beginning of any cheapest timeslot

(c) start at random position in first cheapest timeslot

(d) start at beginning of first cheapest timeslot

(e) increasing price

(f) constant power usage despite of increasing price

Fig. 1 Top: pricing curve; Center: power usage; Bottom: some timeslots and running times

The optimization strategies used by real future devices are unknown at the present time. So we tried various optimization parameters. Figures 1(a)-1(d) show the result of varying two parameters, when the price-curve stayed constant. Any result is optimal for end customers in households. The load is heavily dependent on the optimization strategy used by the devices, consequently there is no simple relationship between price and load.

Another unwanted effect is shown in Figures 1(f) and 1(e). The power usage might remain (nearly) constant even though the price is rising.

3 Combined Heat and Power Plants (CHPP) for Load Balancing in a Virtual Power Plant

The virtual power plant (VPP) mainly consists of three power-generating devices. These are solar power, wind power and combined heat and power plants. The VPP administrator faces the difficult task of delivering a prespecified amount of power to a provider. The VPP generally commits itself 24 hours in advance and is bound by its forecast. If the VPP deviates from its forecast if faces penalties, i. e. lowered revenue. It is therefore of utmost interest for the VPP to stick to the pre-announced amounts of deliverable energy. Yet, sun and wind can change unpredictably. The duty of the VPP administrator is now to put the CHPP in the VPP to good use. The administrator can either switch on more CHPPs to generate more power. Or the administrator can switch them off thereby accommodating an increased amount of power produced by solar or wind power. In any case it is important to know the undisturbed amount of power produced by the CHPPs.

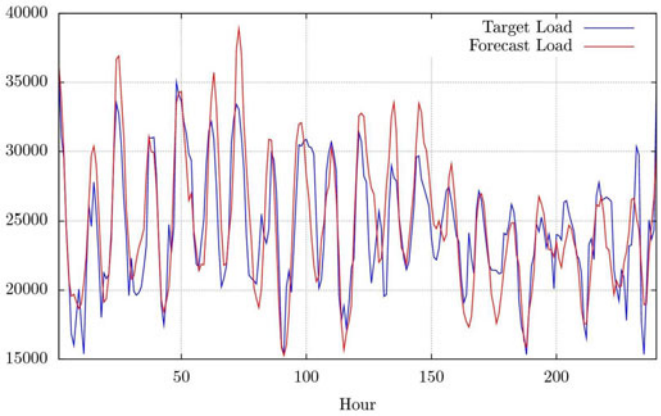


Fig. 2 24-hour neural network forecast of CHPP load (in kWh). The VPP administrator is especially interested in getting the turning points right. This is generally achieved by our forecast.

In our approach we forecast the power produced by CHPPs with neural networks using a standard 3-layer perceptron. As the electric power produced by CHPPs is merely a by-product of heat generation we start our analysis with weather data at hourly intervals. We generate forecasts for up to 24 hours in advance. Our initial model includes the following data: 48 hours CHPP load, 48 hours of outside temperature, 24 hours of weather forecasts for outside temperature, 48 hours of global radiation, 48 hours of wind speed. The training dataset ranges from October 1st, 2005 until September, 30th, 2007. The validation dataset includes the timespan April, 1st, 2007 until September, 30th, 2007.

We successively reduce the number of inputs watching the error on the validation dataset. It turns out that the only time series really needed for achieving accurate

forecasts is the outside temperature. The final model retained contains 48 hours of past outside temperature and 24 hours of weather forecasts. Figure 2 shows the results. It is important to note, that the forecasts gets the turning points right: that is, the time when load will start to increase or decrease is correctly forecast. This is very relevant for the VPP administrator because it allows for load balancing by switching the CHPPs on or off accordingly if needed.

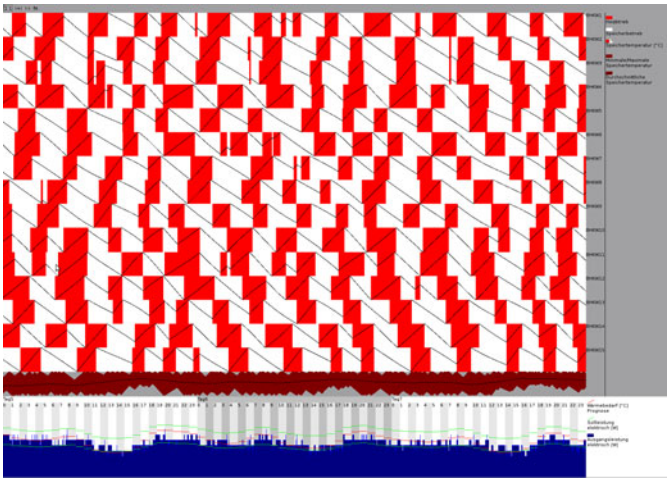


Fig. 3 Simulation of 15 CHPPs, centrally controlled to fit the prespecified amount of power. Top: storage temperature and state of each CHPP; Center: Aggregated storage temperatures; Bottom: Prognosted heat requirement, prespecified and realized power output.

Our software prototype creates a number of independent CHPPs, with randomly initialized storage temperature, running state and an annual operation time between 3000 and 5000 hours. The characteristics of the simulated devices are based on "Dachs"-series CHPPs of the european market leader SenerTec. To allow load balancing a 750l heat water storage is attached to the CHPP. The storage temperature is bounded above and below.

We use the results from neural network forecast of heat requirement and a prespecified profile (minimum/maximum) for electrical power output as input data for the simulated CHPPs and the central CHPP-controller. Several controller strategies were implemented. Figure 3 shows a controller which observes the power output of CHPPs and their running times and switches CHPPs on/off when the output will under-/overrun the target range. This strategy allows a minimal number of switching operations. Optimizing to an accurate target output value would result in a very high switching frequency. Zero power output or maximal power output (CHPPs off/running at the same time) is possible for short-time, but long-time divergence between heat requirement and prespecified power output will result in unwanted peaks in power output.

4 Conclusions and Outlook

Our work analyzes the inclusion of regenerative energies in a future power system. It shows that a significant increase of the regenerative proportion is possible using adequate control technology on both the consumer and the producer side. However there are also several question marks remaining concerning the actual state of the power markets in 2020 — the target date set by the German government for having technologies in place to achieve at least a 20 percent proportion of regenerative energies in production. It is still unclear how exactly the appliances on the consumer side will behave. Also, totally flexible power rates will probably not be accepted by the regulatory authorities and are arguably not in the best interest of consumers. Concerning the producers there is only limited experience with VPPs in general and no experience with automatic control of CHPPs. A regulatory framework has still to be established.

References

1. Sharon Betz and H. Vincent Poor. Energy Efficiency in Multi-hop CDMA Networks: A Game Theoretic Analysis. 2006.
2. Philipp Brandt. IT in der Energiewirtschaft. *Wirtschaftsinformatik*, 49(5): 380–385, October 2007.
3. Geert Deconinck. Metering, Intelligent Enough for Smart Grids? In *Securing Electricity Supply in the Cyber Age*, volume 15 of *Topics in Safety, Risk, Reliability and Quality*, pages 143–157. Springer Netherlands, 2010.
4. J. Dibbern, T. Goles, R. Hirschheim, and B. Jayatilaka. Information systems outsourcing: a survey and analysis of the literature. *The DATA BASE for Advances in Information Systems*, 35(4): 6–102, 2004.
5. Fettke. State of the Art of the State of the Art – A study of the research method "Review" in the information systems discipline. *Wirtschaftsinformatik*, 48(4): 257–266, 2006.
6. H. Frey. Lastmanagement mit intelligenten Tarifen. *e & i Elektrotechnik und Informationstechnik*, 126(10): 358–364, October 2009.
7. Dirk Hinke, Eva Marie Kurscheid, and Margarit Miluchev. Wirtschaftlichkeitsanalyse eines virtuellen Minutenreserve-Kraftwerks aus dezentralen Klein-Kraft-Wärme-Kopplungsanlagen. *Zeitschrift für Energiewirtschaft*, 33(2): 127–134, June 2009.
8. Andreas Kamper and Anke Eßer. Strategies for Decentralised Balancing Power. In *Biologically-Inspired Optimisation Methods*, volume 210 of *Studies in Computational Intelligence*, pages 261–289. Springer Berlin/Heidelberg, 2009.
9. D. Nestle, J. Ringelstein, and P. Selzam. Integration dezentraler und erneuerbarer Energien durch variable Strompreise im liberalisierten Energiemarkt. *uwf – UmweltWirtschaftsForum*, 17(4): 361–365, 2009.
10. Richard P. O'Neill, Emily Bartholomew Fisher, Benjamin F. Hobbs, and Ross Baldick. Towards a complete real-time electricity market design. *Journal of Regulatory Economics*, 34(3): 220–250, 2008.
11. E.-B. Swanson and N.-C. Ramiller. Information systems research thematics: Submissions to a new journal, 1987–1992. *Information Systems Research*, 4(4): 299–330, 1993.
12. Serafin von Roon and Michael Steck. Dezentrale Bereitstellung von Strom und Wärme mit Mikro-KWK-Anlagen. *uwf – UmweltWirtschaftsForum*, 17(4): 313–319, 2009.
13. B. Wille-Hausmann, T. Erge, and C. Wittwer. Decentralised optimisation of cogeneration in virtual power plants. 2010.

Stationary or Instationary Spreads – Impacts on Optimal Investments in Electricity Generation Portfolios

Katrin Schmitz, Christoph Weber, and Daniel Ziegler

Abstract It is common practice to base investment decisions on price projections which are gained from simulations using price processes. Particularly in the electricity industry with its diverse fuels, homogenous output and long-lived assets the choice of the underlying process is crucial for the simulation outcome. At the most fundamental level stands the question of the existence of stable long-term cointegration relations. Since this question is very difficult to answer empirically, it is also appropriate to investigate the implications of varying assumptions. Not only the specific parameter values but also the specification of the price processes has therefore to be scrutinized. In the presence of fuel price risk, portfolio diversification will usually be drawn into consideration in investment decisions. Therefore we examine the impacts of different ways to model price movements in a portfolio selection model for the German electricity market. Three different approaches of modelling fuel prices are compared: Initially, all prices are modelled as correlated random walks. Thereafter the coal price is modelled as random walk and leading price while the other prices follow through mean-reversion processes. Last, all prices are modelled as mean reversion processes with correlated residuals. The prices of electricity base and peak futures are simulated using historical correlations with gas and coal prices. Yearly base and peak prices are transformed into an estimated price duration curve followed by the steps power plant dispatch, operational margin and NPV calculation and finally the portfolio selection. The analysis shows that the chosen price process

Katrin Schmitz
Chair for Management Science and Energy Economics, Duisburg-Essen University, 45117 Essen, Germany e-mail: katrin.schmitz@uni-due.de

Christoph Weber
Chair for Management Science and Energy Economics, Duisburg-Essen University, 45117 Essen, Germany e-mail: christoph.weber@uni-due.de

Daniel Ziegler
Chair for Management Science and Energy Economics, Duisburg-Essen University, 45117 Essen, Germany e-mail: daniel.ziegler@uni-due.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_83,

assumptions have significant impacts on the resulting portfolio structure and the weights of individual technologies.

1 Introduction

Among the many different areas of interest which have been in the focus of energy-economical research in the last years, investment planning naturally takes a central role. On the one hand investment planning issues are of high relevance both from a social welfare and from a single investor's perspective simply through their high capital intensity. On the other hand, their complexity gives much room for applying a wide range of methodical tools and instruments. This very complexity usually forces investors to mask out the majority of influences which are relevant for the decision or to take them in only in a strongly simplified form. The homogeneity of the product electricity implies that the profitability of an investment in generation capacity (in the case of classic thermal technologies) can almost exclusively be characterized as a function of the development of electricity, fuel and CO_2 emission prices. From finance there comes the established (and, all in all, successful) practice to treat price developments as stochastic processes. The fundamental drivers of these processes are, if at all, only drawn into consideration in an intermediate way as the determinants of the parameters which rule these processes (unless they are simply taken from historical samples). This practice has also found widespread adoption in energy economics. The selection of suitable processes and their adequate parametrization thus determines the outcome and the quality of an investment decision exclusively. The difficulty of parametrization is immediately evident and does not need further discussion. More interesting and, in a way, more alarming is that impacts of the underlying stochastic processes are often not clear at all.

Therefore, our study aims at examining the sensitivity of an investor's decision to his choice of price process. We consider several different processes and use them to simulate the returns of different investment options in a multi-stage process. Then, we analyze which optimum investment strategies would arise for investors with varying risk aversion. In particular, we measure the scale of variations between the different model approaches.

The question of process selection in the context of energy markets has been addressed in [4]. The principles of portfolio selection have been formulated in [7]. While most works focus on the social welfare perspective (e.g. in [1], [10] or [6]) the work of [8] is, to our knowledge, the only recent publication which takes a single

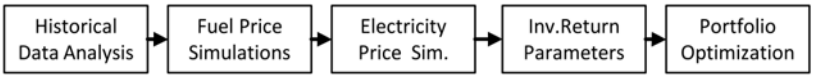


Fig. 1 Model Overview

investor's perspective[1]. There the authors refrained from simulating price processes, but performed Monte Carlo simulations directly based on the historical price distributions.

2 Model Description

The principle course of actions is sketched in figure 1. On the base of one single set of historical price data, we estimate the required parameters of all price processes. These are used for simulating annual mean values over a long period of time for electricity and fuel prices. To be more precise: We directly simulate the variable costs of production for our set of generation technologies. In the next step we produce expected price duration curves for each year. Using current technology data, we then simulate a unit commitment and calculate the corresponding cash flows of each investment project. Afterwards, we calculate the present values through discounting and determine optimum investment portfolios for different levels of risk aversion.

In our study we focus on investments in the dominating thermal generation technologies nuclear, lignite, hard coal, gas turbine and CCGT. In the following we refer to these as NUC, LIG, COA, GAS and CCG. We calculate the historic volatilities from data provided by ([9]). The mean values of the future fuel price developments are based on the forecasts of the IEA (cp. [3]). The parameters of the generation technologies are given by [5].

Table 1 Fuel Price Process Modelling (with $x_{tec} = ln(c_{var,tec,t})$)

	COA	$tec \in GAS, CCG$
A	$\Delta x_{COA,t} = \varepsilon_{COA,t} \cdot \sigma_{hist,COA} + \mu_{corr,COA}$	$\Delta x_{tec,t} = \varepsilon_{tec,t} \cdot \sigma_{hist,tec} + \mu_{corr,tec}$
B	$\Delta x_{COA,t} = \varepsilon_{COA,t} \cdot \sigma_{hist,COA} + \mu_{corr,COA}$	$\Delta x_{tec,t} = \kappa_{tec}(\theta_{tec,t} + x_{COA,t} - x_{tec,t}) + \sigma_{tec} \cdot \varepsilon_{tec,t}$
C	$\Delta x_{COA,t} = \kappa_{COA}(\mu_{COA,t} - x_{COA,t}) + \sigma_{COA} \cdot \varepsilon_{COA,t}$	$\Delta x_{tec,t} = \kappa_{tec}(\theta_{tec,t} + x_{COA,t} - x_{tec,t}) + \sigma_{tec} \cdot \varepsilon_{tec,t}$

We simulate the variable generation costs in the following variants: LIG is assumed to be constant in all variants in real terms since fuel costs are production costs and not driven by market prices. Due the the insufficient historical data, CO_2 emission prices are not modelled stochastically in our study. Instead, we have set a fix price. For a real investment decision, this would be, of course, an ineligible simplification. However, against the background of the limited purpose of this work, this simple handling appears acceptable. We have decided to model NUC as independent random walk in all variants since the historical evidence for interdependencies with

[1] The standard finance theory states that only investors should diversify, not companies. Yet in the presence of insolvency costs a diversification of generation companies can be advantageous.

other fuel prices are very weak. (In addition, the variable costs of NUC consist to a considerable extent of costs for storage, transportation, postprocessing and discounted future disposal costs which are hard to forecast but are surely independent of other fuel prices.) In variant A the differences of the production cost logs of COA, GAS and CCG are modelled as correlated random walks. In the variants B and C COA is assumed to be price-leading. The spread between COA and the natural gas based technologies is modelled as a mean reverting process. While in B the COA production costs themselves are still modelled as a random walk, they too follow a mean reversion process in C. All modelling approaches are summarized in table 1[2].

A sound estimation of the electricity base and peak price parameters is difficult given that we cannot use price data prior to the market liberalization. Hence, the set of price data which can be used is limited to a relatively short period (cp. [2]). Due to the long-term mean reverting character of electricity prices (by reason of shutdowns and constructions of new power plants) and the interdependency with fuel prices the electricity prices are modelled by the relationship[3]: $\Delta_{k,t} = \beta_{0,k}(p_{k,t-1} - \beta_{b,k}k_{var,b,t-1} - \beta_{p,k}k_{var,p,t-1}) + \varepsilon_{k,t}\sigma_k, \quad \varepsilon_{k,t} \sim N(0,1), k \in B,P$

Due to the non-storability of electricity the hourly electricity prices are important for calculating the operation margins. The distribution of hourly electricity prices within one year is approximated by a price duration curve. We model this by means of an exponential function which is parameterized with the simulated base and peak prices. Following the definition of the Phelix Base Future we assume that yearly average equals this Base Future value. The Phelix Peak value is used as a proxy for the mean of the 3000 hours with the highest prices. Consequently, the price duration curve is described by the function: $p_{h,t} = a_t \cdot h^{-\beta_t}$ The parameters α_t and β_t are calculated by: $\beta_t = 1 - log_{8760/3000}(\frac{8760 p_{B,t}}{3000 p_{P,t}}), \quad \alpha_t = 3000 \cdot p_{P,t} \cdot (1-\beta_t) \cdot 3000^{-(1-\beta_t)}$

The unit commitment is performed as a simple comparison between electricity price in each hour and the variable generation costs in the respective year. If the resulting operational margin is positive, it is assumed that the unit operates at full capacity; if not, it is assumed to be offline. The annual operational margin simply is the sum over the hourly values. This approach implies that we ignore several restrictions in power plant operation, in particular minimum and maximum online and offline times, startup costs etc. However, we are convinced that these play a minor role within our focus. For sure, this is a field for potential improvement if one intended to use this approach or a similar one for an investment decision. By discounting all

[2] $\Delta x_{tec,t}$ = difference of price logs t and t-1; $\varepsilon_{tec,t}$ = correlated standard normally-distributed random variable; $\sigma_{tec,t}$ = standard deviation of price log difference; $\mu_{corr,tec}$ = adjusted mean of price log difference; $\kappa_{tec,t}$ = mean reversion rate; $\theta_{tec,t}$ = equilibrium spread between fuel and coal

[3] $k \in \{B,P\}$ = index of electricity product:B = Phelix Base, P = Phelix Peak; $\beta_{0,k}$ = reversion speed towards equilibrium relation; $\beta_{b,k}k_{var,b,t-1}$ = relative equilibrium level of the baseload technology $(= min\{x_{COA,t}, x_{CCG,t}\})$; $\beta_{p,k}k_{var,p,t-1}$ = relative equilibrium level of the peakload technology (gas)

operational margins, annual fix and investment costs to the present, we get the capital value of the investment project in the usual way. Over all simulation runs we thus get the multi-variate distribution of investment values and its distribution parameters mean values $\mu_{inv,tec}$, variances and covariances $\sigma_{tec,tec'}$. An optimum investment strategy can be determined with the classic Markowitz approach by maximizing the utility as given by [4] $max_w \mu_{npv}^T \cdot w - 0.5 \cdot A \cdot w^T \Sigma w \quad s.t. \quad w^T \cdot 1 = 1$

Table 2 Results - Specific NPVs - means and standard deviations

	μ_{COA}	μ_{LIG}	μ_{GAS}	μ_{CCG}	μ_{NUC}		σ_{COA}	σ_{LIG}	σ_{GAS}	σ_{CCG}	σ_{NUC}
A	-0.06	0.08	-0.37	0.36	0.55	A	0.61	0.69	0.33	0.39	0.54
B	-0.01	0.13	-0.50	0.11	0.59	B	0.29	0.56	0.14	0.26	0.41
C	0.02	0.13	-0.51	0.10	0.59	C	0.21	0.37	0.13	0.25	0.24

3 Results

Our results can be summarized as follows:

- Table 2 shows the expected values for the specific NPVs of the different investment options. The values show big differences between the different options. The ranking does not change significantly over the variants A, B and C.
- Both lignite and nuclear can be classified as high risk technologies (except for NUC in variant C). This meets intuition and also fits well to the results of [8]: Both technologies have stable variable costs. Consequently, the fluctuations of

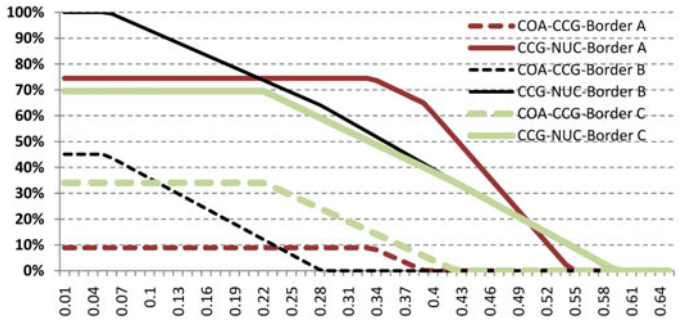


Fig. 2 Technology Weights for different levels of required expected specific NPVs

[4] w = vector of relative weights of investments; μ_{npv} = vector of expected specific NPVs of all investment options; Σ = covariance matrix of specific NPVs; A = risk aversion parameter

electricity prices lead to corresponding fluctuations in their returns. In contrast, in the case of gas and coal technologies, one can observe stable spreads between costs and therefore more stable operational margins. In variants A and B, CCGT is the technology with the lowest ROI volatility and therefore plays a major role in portfolios with a high risk aversion. Surprisingly, NUC reaches an even lower ROI volatility in variant C.

- NUC dominates LIG over the whole space of possible risk aversion levels.
- As one would expect, one can observe that ROI volatility of the technologies with stable fuel costs, NUC and LIG, is reduced to a similar extent by exchanging the random walk of the gas price process (variant A to B) and of the coal price process (variant B to C). In contrast, the main decrease of ROI volatility of COA, CCG and GAS is reached through switching from the random walk of the gas price to a mean reversion process while the change in the coal price process has no significant impact.
- Figure 2 describes the resulting portfolio structures.[5] A high risk aversion results in mixed portfolios consisting of CCG, COA and, in C, of NUC. With increasing risk tolerance, the share of NUC in the portfolio rises and replaces CCG and COA until the portfolio consists only of a pure NUC investment.

References

1. S. Awerbuch and M. Berger. Applying portfolio theory to EU electricity planning and policy-making. Report number EET/2003/03, IEA, 2003.
2. EEX. Phelix baseload year future, phelix peakload year future, 2009.
3. IEA. World energy outlook 2008. International Energy Association (IEA), 2008.
4. Blake Johnson and Graydon Barz. *Energy modelling – Advances in the Management of Uncertainty*. chapter Selecting stochastic processes for modelling electricity prices, pages 9–58. Risk Books, 2005.
5. P. Konstantin. *Praxisbuch Energiewirtschaft: Energieumwandlung, -transport und -beschaffung im liberalisierten Markt*. Springer, 2nd edition, 2009.
6. Boris Krey and Peter Zweifel. Efficient electricity portfolios for switzerland and the united states. Working Papers 0602, University of Zurich, Socioeconomic Institute, February 2006.
7. Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1): 77–91, 1952.
8. F.A. Roques, D.M. Newbery, and W.J. Nuttall. Fuel mix diversification incentives in liberalized electricity markets: A mean-variance portfolio theory approach. *Energy Economics*, 30(4): 1831–1849, 2008.
9. Sachverständigenrat. Energy prices in Germany. Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (German Council of Economic Experts), 2008.
10. M. Sunderkötter and C. Weber. Valuing fuel diversification in optimal investment policies for electricity generation portfolios. EWL Working paper 06/09, 2009.

[5] For given level of required specific portfolio NPV: share of COA = difference between the x-axis and the respective dotted line; CCG =distance between dotted and solid line;NUC = distance between solid line and 100percent level

Market Modeling, Forecasting and Risk Analysis with Historical Consistent Neural Networks

Hans-Georg Zimmermann, Ralph Grothmann, Christoph Tietz, and Holger von Jouanne-Diedrich

1 Introduction

Business management requires precise forecasts in order to enhance the quality of planning throughout the value chain. Furthermore, the uncertainty in forecasting has to be taken into account.

Neural networks (NN) offer significant benefits for dealing with the typical challenges associated with forecasting. With their universal approximation properties, NN make it possible to describe non-linear relationships between a large number of factors and multiple time scales [3]. In contrast, conventional econometrics (such as ARMA, ARIMA, ARMAX) remain confined to linear systems [8]. A wide range of models is discussed within the class of neural networks. For example, in terms of the data flow, it is possible to draw a distinction between feedforward and (time) recurrent NNs [1]. In this paper we focus on recurrent NN.

It is noteworthy that any NN equation can be expressed as an architecture which represents the individual layers in the form of nodes and the connections between the layers in the form of links. This relationship will be described hereinafter as the correspondence principle between equations, architectures and the local algorithms associated with them. The use of local algorithms provides an elegant basis for the expansion of the NN towards the modeling of large systems.

In this article, we present a new type of recurrent NN, called *historical consistent neural network* (HCNN). HCNNs allow the modeling of highly-interacting non-linear dynamical systems across multiple time scales. HCNNs do not draw any distinction between inputs and outputs, but model observables embedded in the dynamics of a large state space.

In the following, Sec. 2 is dedicated to the theoretical foundations of HCNNs. Sec. 3 reports on the application of HCNN to analyze the risk in financial markets, while sec. 4 summarizes the primary findings and points to practical applications.

Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 Munich, e-mail: Hans_Georg.Zimmermann@siemens.com

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 531
DOI 10.1007/978-3-642-20009-0_84,

2 Historical Consistent Neural Networks (HCNN)

2.1 Modeling Open Dynamic Systems with RNN

To derive the HCNN, let us first consider a simple recurrent neural network (RNN). Our guideline for the development is the correspondence principle of NN (see sec. 1). We start from the assumption that a time series y_τ is created by an open dynamic, which can be described in discrete time τ using a state transition and output equation [3]:

$$\text{state transition } s_\tau = f(s_{\tau-1}, u_\tau) \tag{1}$$

$$\text{output equation } y_\tau = g(s_\tau) \tag{2}$$

$$\text{system identification } E = \frac{1}{T}\sum_{\tau=1}^{T}\left(y_\tau - y_\tau^d\right)^2 \rightarrow \min_{f,g} \tag{3}$$

The time-recurring state transition equation $s_\tau = f(s_{\tau-1}, u_\tau)$ describes the current state s_τ dependent on the previous system state $s_{\tau-1}$ and the external influences u_τ.

Without loss of generality we can approximate the state space model with a RNN [7, 11]:

$$\text{state transition } s_\tau = \tanh(As_{\tau-1} + Bu_\tau) \tag{4}$$

$$\text{output equation } y_\tau = Cs_\tau \tag{5}$$

To cover the complexity of the original task, the RNN has to incorporate an increased state dimension. We use the technique of finite unfolding in time [3] to transform the temporal equations into a spatial architecture (see Fig. 1).

The training of the RNN can be conducted using the error-back-propagation-through-time algorithm [3]. The underlying idea here is that any RNN can be reformulated into an equivalent feedforward neural network, if matrices A, B and C are identical in the unfolded architecture (shared weights). For algorithmic solution methods, the reader is referred to the overview article by B. Pearlmatter [6].

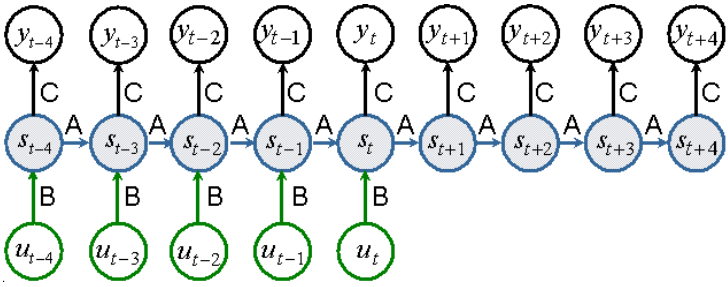


Fig. 1 Unfolded in time recurrent neural network (RNN)[11]. Note, that the hidden state clusters s_τ are equipped with a tangent hyperbolic $\tanh()$ activation function.

2.2 Modeling of Closed Dynamic Systems with HCNN

The RNN is used to model and forecast an open dynamic system using a non-linear regression approach. Many real-world technical and economic applications must however be seen in the context of large systems in which various (non-linear) dynamics interact with each other in time. Projected on a model, this means that we do not differentiate between inputs and outputs but speak about observables. Due to the partial observability of large systems, we need hidden states to be able to explain the dynamics of the observables. Observables and hidden variables should be treated by the model in the same manner. The term observables embraces the input and output variables (i. e. $Y_\tau := (y_\tau, u_\tau)$). If we are able to implement a model in which the dynamics of all of the observables can be described, we will be in a position to close the open system.

Motivated by these modeling principles, we develop the HCNN as follows:

$$\text{state transition } s_\tau = \tanh(As_{\tau-1}) \tag{6}$$

$$\text{output equation } Y_\tau = [Id, 0]s_\tau \tag{7}$$

$$\text{system identification } E = \sum_{\tau=t-m}^{t} \left(Y_\tau - Y_\tau^d\right)^2 \to \min_A \tag{8}$$

The HCNN (Eq. 6 and 7) describes the dynamics of all observables by the sequence of states s_τ using a single transition matrix A. The observables ($i = 1, \ldots, N$) are arranged as the first N neurons of a state s_τ, whereas the hidden variables are represented by the subsequent neurons. The connector $[Id, 0]$ is a fixed matrix of appropriate size which reads out the observables. The HCNN is unfolded across the entire time path, i. e. we learn the unique history of the system. The HCNN architecture is depicted in Fig. 2.

Teacher Forcing (TF) [10, 6] was originally introduced as an extension to the algorithmic training procedures of RNNs. In contrast, we formulate TF as a part of the NN architecture, which allows us to learn the NN using standard error-backpropagation-through-time. Fig. 3 deals with the resulting HCNN architecture.

Let us explain the TF mechanism in the extended HCNN architecture (see Fig. 3): In every time step $\tau \leq t$ the expected values for all observables Y_τ are replaced by the observations Y_τ^d using an intermediate (hidden) layer r_τ. Up to present time ($\tau = t$), the output layers hold the difference between expectations Y_τ and observations Y_τ^d. The output layers are given fixed target values of zero. This causes the HCNN to learn the expectations Y_τ to compensate for the observations $-Y_\tau^d$.

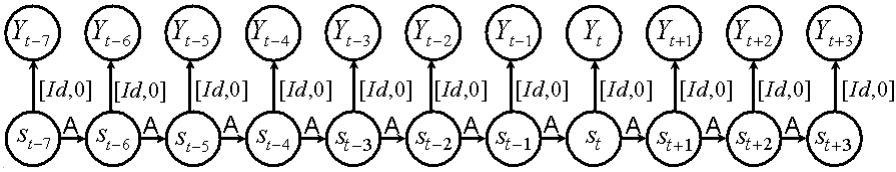


Fig. 2 Architecture of the Historical Consistent Neural Network (HCNN)

The content of the output layer ($Y_\tau - Y_\tau^d$), is negated and transferred to the first elements of r_τ using the fixed $[-Id,0]'$ connector. In addition, the expected values Y_τ are transferred from the internal state s_τ to r_τ. The net effect is that the expected values Y_τ on the first N components of the state s_τ are replaced by the corresponding observations Y_τ^d (i. e. $r_{\tau=1,\ldots,N} = Y_\tau - (Y_\tau - Y_\tau^d) = Y_\tau^d$, see Fig. 3). Since all additional connectors for the mechanism introduced are fixed and are used only to transfer data in the network, TF does not lead to a larger number of free network parameters.

In the future part the HCNN is iterated exclusively on the basis of expected values. This turns an open system into a closed dynamical system and we do not need to make the assumption of an constant environment as for dynamical systems. The usage of TF does not reintroduce an input / output modeling, since we replace the expected value of the observables with their actual observations. For sufficiently large HCNNs and convergence of the output error to zero, this architecture converges towards the fundamental HCNN architecture shown in Fig. 2.

Eq. 9 to 11 show the state transition and output equation derived from the HCNN architecture depicted in Fig. 3.

$$\text{state transition } \forall \tau \leq t \; s_\tau = \tanh\left(A\left(s_{\tau-1} - [Id,0]^T\left(Y_\tau - Y_\tau^d\right)\right)\right) \tag{9}$$

$$\forall \tau > t \; s_\tau = \tanh\left(A s_{\tau-1}\right) \tag{10}$$

$$\text{output equation } \forall \tau \in t \; Y_\tau = [Id,0] s_\tau \tag{11}$$

The HCNN models the dynamics of all observables and hidden variables in parallel. Thus, a high-dimensional state transition matrix A is required. For numerical stability we recommend to initialize matrix A with a (random) sparse connectivity. The degree of sparsity is chosen with respect to the metaparameters connectivity and memory length (for details see [12]).

If we repeat the system identification we will get an ensemble of solutions. All solutions have a model error of zero in the past, but show a different behavior in the future. The reason for this lies in different ways of reconstructing the hidden variables from the observations and is independent of different random sparse initializations. Since every model gives a perfect description of the observed data, we can use the simple average of the individual forecasts as the expected value, assuming that the distribution of the ensemble is unimodal.

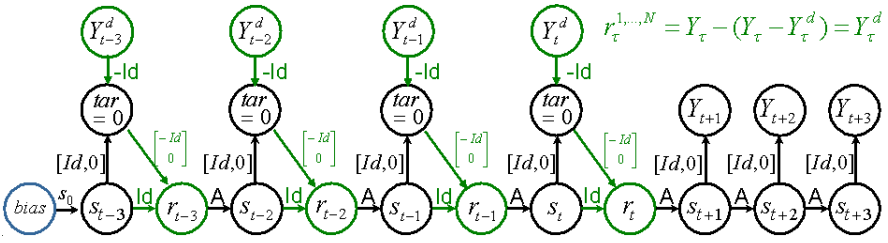


Fig. 3 HCNN incorporating a Teacher Forcing (TF) mechanism

3 Risk Analysis in Financial Markets

The latest financial crisis has triggered a far-reaching discussion on the limitations of quantitative forecasting models and made investors very conscious of risk [2]. Risk management frequently considers the probability distribution of market prices / returns [4]. In order to understand risk distributions, traditional risk management uses historical simulations which require strong model assumptions. Risk is understood as a random walk, in which the diffusion process is calibrated by the observed past error of the underlying model [5].

For our approach this concept fails, because the (past) residual error of the HCNN is zero. Our risk concept is based on the partial observability of the world, leading to different reconstructions of the hidden variables and thus, different future scenarios. Since all scenarios are perfectly consistent with the history, we do not know which of the scenarios describes the future trend best and risk emerges.

Our approach directly addresses the model risk. For HCNN modeling we claim that the model risk is equal to the forecast risk. The reasons can be summarized as follows: First, HCNNs are universal approximators, which are therefore able to describe every future market scenario. Second, the form of the ensemble distribution is caused by underlying dynamical equations, which interpret the market dynamics as the result of interacting decisions [13]. Third, in experiments we have shown that the ensemble distribution is independent from the details of the model configuration, iff we use large models and large ensembles.

The diagram below (Fig. 4, left) shows our approach applied to the Dow Jones Industrial Index (DJX). For the ensemble, a HCNN was used to generate 250 individual forecasts for the DJX.

For every forecast date, all of the individual forecasts for the ensemble represent the empirical density function, i. e. a probability distribution over many possible market prices at a single point in time (see Fig. 4, right). It is noticeable that the actual development of the DJX is always within the ensemble channel (see grey lines, Fig. 4, left). The expected value for the forecast distribution is also an adequate point forecast for the DJX (see Fig. 4, right).

It is our intention to use the ensemble distribution of the HCNN in a future project for the evaluation of call and put options.

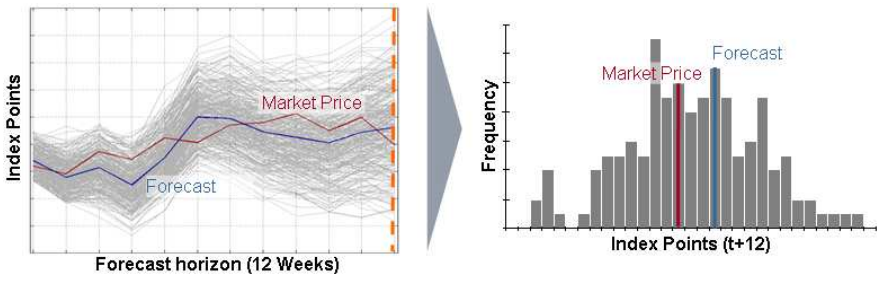


Fig. 4 HCNN ensemble forecast for the Dow Jones Index (12 week forecast horizon), left, and associated index point distribution for the ensemble in forecast time step $t + 12$ weeks, right.

4 Conclusion and Outlook

The joint modeling of hidden and observed variables in large recurrent neural networks provides new prospects for planning and risk management. The ensemble approach based on HCNN offers an alternative approach to forecasting of future probability distributions. HCNNs give a perfect description of the dynamic of the observables in the past. However, the partial observability of the world results in a non-unique reconstruction of the hidden variables and thus, different future scenarios. Since the genuine development of the dynamic is unknown and all paths have the same probability, the average of the ensemble may be regarded as the best forecast, whereas the bandwidth of the distribution describes the market risk.

Today, we use HCNN forecasts to predict prices for energy and precious metals to optimize the timing of procurement decisions. Work currently in progress concerns the analysis of the properties of the ensemble and the implementation of these concepts in practical risk management and financial market applications.

All NN architectures and algorithms are implemented in the Simulation Environment for Neural Networks (SENN), a product of Siemens Corporate Technology.

References

1. Calvert D. and Kremer St.: Networks with Adaptive State Transitions, in: Kolen J. F. and Kremer, St. (Ed.): A Field Guide to Dynamical Recurrent Networks, IEEE, 2001, pp. 15–25.
2. Föllmer, H.: Alles richtig und trotzdem falsch?, Anmerkungen zur Finanzkrise und Finanzmathematik, in: MDMV 17/2009, pp. 148–154
3. Haykin S.: Neural Networks and Learning Machines, 3rd Edition, Prentice Hall, 2008.
4. Hull, J.: Options, Futures & Other Derivative Securities. Prentice Hall, 2001.
5. McNeil, A., Frey, R. and Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools, Princeton University Press, Princeton, New Jersey, 2005.
6. Pearlmatter B.: Gradient Calculations for Dynamic Recurrent Neural Networks, in: A Field Guide to Dynamical Recurrent Networks, Kolen, J.F.; Kremer, St. (Ed.); IEEE Press, 2001, pp. 179–206.
7. Schäfer, A. M. und Zimmermann, H.-G.: Recurrent Neural Networks Are Universal Approximators. ICANN, Vol. 1., 2006, pp. 632–640.
8. Wei W. S.: Time Series Analysis: Univariate and Multivariate Methods, Addison-Wesley Publishing Company, N.Y., 1990.
9. Werbos P. J.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD Thesis, Harvard University, 1974.
10. Williams R. J. and Zipser, D.: A Learning Algorithm for continually running fully recurrent neural networks, Neural Computation, Vol. 1, No. 2, 1989, pp. 270–280.
11. Zimmermann, H. G., Grothmann, R. and Neuneier, R.: Modeling of Dynamical Systems by Error Correction Neural Networks. In: Soofi, A. und Cao, L. (Ed.): Modeling and Forecasting Financial Data, Techniques of Nonlinear Dynamics, Kluwer, 2002.
12. Zimmermann, H. G., Grothmann, R., Schäfer, A. M. and Tietz, Ch.: Modeling Large Dynamical Systems with Dynamical Consistent Neural Networks, in New Directions in Statistical Signal Processing: From systems to brain. Haykin, S., Principe, J. C., Sejnowski, T. J., and McWhirter, J. (Ed.), MIT Press, Cambridge, Mass., 2006.
13. Zimmermann, H. G.: Neuronale Netze als Entscheidungskalkül. In: Rehkugler, H. und Zimmermann, H. G. (Ed.): Neuronale Netze in der Ökonomie, Grundlagen und wissenschaftliche Anwendungen, Vahlen, Munich 1994.

III.3 Health Care Management

Chair: Prof. Dr. Teresa Melo (Hochschule für Technik und Wirtschaft des Saarlandes (HTW))

The field of health care provides a broad range of applications suitable for analysis using Operations Research (OR) techniques. Recent technological innovations in medicine and health care as well as advances in information and communication technologies further broadened the application scope of OR models and methods in health care. In addition to cost savings, health care OR improves efficiency and effectiveness in health care service delivery.

Submissions are welcomed that involve innovative applications of existing OR models and/or the development of new models and techniques. Topics of interest include, but are not restricted to, management and optimization of health care operations, resource allocation, capacity planning, workforce staffing, treatment design, patient flow modeling, quality and performance measurement, disease prevention modeling.

Dynamic Simulations of Kidney Exchanges

M. Beccuti, V. Fragnelli, G. Franceschinis, and S. Villa

Abstract In this paper we develop a simulator modeling a kidney exchange program, in which donor-recipient pairs with characteristics drawn from a distribution based on real data join the system over time, and a centralized authority organizes a suitably chosen set of exchanges among the pairs in the pool at regular intervals of time, as it happens in the Netherlands or in the US. We compare and discuss the results of numerical simulations on this model varying the matching policy.

1 Introduction

Kidney transplantation is the elective treatment for patients with irreversible kidney failure. Recently, kidney exchanges are emerging as a viable option for utilizing live donor kidneys from those who are incompatible with their intended recipients. Such exchanges involve two or more incompatible pairs for which a reciprocal compatibility holds, and were first suggested by the medical community, allowing to increase the number of transplants [7]. Nowadays kidney exchange programs (KEP) have been started and are currently going on in many countries, the leading examples being the US and the Netherlands, and including Germany, New Zealand, Australia, Italy among other [3] [4], [6].

The organization of KEP raises several questions that can be addressed from a mathematical perspective, and, starting from [8], a wide literature is devoted to the study of the optimal organization of such exchanges in a static situation, see [9],

M. Beccuti
Università degli Studi di Torino, Torino, Italy, e-mail: beccuti@di.unito.it

V. Fragnelli and G. Franceschinis
Università degli Studi del Piemonte Orientale, Alessandria, Italy, e-mail: { vito.fragnelli, giuliana.franceschinis } @mfn.unipmn.it

S. Villa
Università degli Studi di Genova, Genova, Italy, e-mail: villa@dima.unige.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_85,

[5], [2] and references therein. Considerably less attention has been devoted to the dynamic setting, though the latter is fairly more realistic than the static model, since donor-recipient pairs join the system over time and not all at the same moment. The main question to be answered is how to find an optimal matching policy, i.e. a suitable scheduling of the exchanges in order to maximize a properly chosen objective.

A theoretical analysis of the dynamic problem is carried out in [11], under some "long run" assumptions. More precisely, it is proved that if only paired exchanges are considered and if there is an unlimited availability of the so called "underdemanded pairs" (pair types that are difficult to match) then the policy maximizing a discounted sum of the total number of exchanges is the one organizing exchanges as soon as they become available.

In this work we develop a simulation tool which models a realistic dynamic situation, in which a centralized authority organizes a suitably chosen set of exchanges among the pairs in the pool at regular intervals of time. The goal is to determine the potential impact of a long term organization of kidney exchanges: we discuss the results of numerical simulations obtained using the proposed model, analyzing the performance of a given exchange policy, in terms of some relevant quantities such as the number of matched and unmatched patients, and the average waiting time w.r.t. the pair characteristics. Moreover, a comparison of the performance corresponding to different choices of the time interval between one slot of exchanges and the next one is presented. Our work is related to the results in [10, 1], where an experimental analysis of different matching policies is presented. In particular, for evaluating the policy, we account the average waiting time for pairs in the system, an aspect which is not considered in [1]. On the other hand they propose a (sub)optimal policy that in terms of total number of performed transplants behaves better than ours. A careful study of the waiting times corresponding to a fixed policy belonging to the class described in our work can be found in [10]. Our contribution w.r.t. this paper is the long term perspective and a more extensive comparison among different policies. In [12] a different context is considered, where pairs can choose among paired exchanges and the waiting list from deceased donors. A simplified version of the real situation is taken into account, with only two possible pair types and an optimal decision strategy is proposed.

The paper is organized as follows: we start introducing the static kidney exchange problem together with the compatibility assumptions in Section 2, then we describe the dynamic case and the experimental analysis in Section 3, Section 4 concludes.

2 General Setting

As it happens in the real world, we assume that only incompatible pairs are admitted to KEP. Incompatibility between donor and recipient has two sources: a blood type incompatibility or a tissue type incompatibility (also called positive crossmatch). While the first one is simple to check, the second is difficult to predict, but the probability of positive crossmatch between two unrelated individuals is 11% . The com-

position of the pool for kidney exchange programs is obtained combining the known probability distribution over blood types with the probability of positive crossmatch, according to the model proposed in [12, 10], and is reported in Tab. 1.

Table 1 Blood type characteristics of Simulated Pairs (pair percentage)

Blood Type	Recipient 0		Recipient A		Recipient B		Recipient AB	
Donor 0	[1]	14.0	[2]	6.3	[3]	2.4	[4]	0.5
Donor A	[5]	37.8	[6]	6.8	[7]	6.1	[8]	0.5
Donor B	[9]	12.0	[10]	5.1	[11]	1.2	[12]	0.2
Donor AB	[13]	2.0	[14]	2.8	[15]	2.1	[16]	0.1

We assume that each patient is indifferent among two compatible kidneys and we do not take into account the possibility of having a positive crossmatch when organizing the exchanges, so that subsequent results overestimate the number of possible exchanges. This assumptions, while simplifying the real situation, are considered acceptable also from a medical point of view and are common in the literature studying KEPs [9]. We therefore deal with a set of pair types $T = \{1, \ldots, t\}$ (in our situation $t = 16$ according to the numbers in square brackets in Tab. 1).

We model the compatibility by introducing a non-negative symmetric compatibility matrix R, where $R_{ij} > 0, i, j \in T$ means that an exchange between pairs of type i and j is possible with a *revenue* R_{ij} (e.g. the compatibility level); by the symmetry of R, we can restrict to the upper triangle, i.e. $i \leq j$. Mathematically speaking a *static kidney exchange problem* consists of a compatibility matrix R and a set of pairs, that can be identified with a tuple $(s_1, \ldots, s_t)$ describing the number of pairs of each type. A feasible set of exchanges can be found solving the following integer linear programming problem:

$$\begin{aligned}
\max \quad & \sum_{\substack{i,j \in T \\ \text{s.t. } R_{ij}>0}} R_{ij} x_{ij} \\
\text{s.t.} \quad & \sum_{\substack{i=1,\ldots,k \\ \text{s.t. } R_{ik}>0}} x_{ik} + \sum_{\substack{j=k,\ldots,t \\ \text{s.t. } R_{kj}=1}} x_{kj} \leq s_k \quad k \in T \\
& x_{ij} \in \mathbb{N} \qquad\qquad\qquad\qquad\qquad\qquad i \leq j, \quad i, j \in T
\end{aligned}$$

where $x_{ij}, i \leq j, i, j \in T$ is the number of exchanges involving one pair of type i and one pair of type j. Hereafter, we use a 16×16 matrix where $R_{ij} \in \{0, 1\}$ (pairs are compatible or not); this leads to a solution that maximizes the number of performed transplants.

A similar formulation of a kidney exchange problem is given in [9] using an undirected unweighted graph whose vertices represent the pairs in the system, and the edges connect compatible pairs; they look for a maximum cardinality matching, in order to provide a transplant to as many participants as possible.

The dynamic situation is more complicated and only partial solutions have been proposed so far to tackle it. We describe our approach in the next section.

3 The Dynamic Model: Experimental Analysis

We start this section by describing the dynamic model we simulate, under the compatibility assumptions described above. We assume that each pair enters in the KEP at a certain time, and leaves the system only when receiving a kidney. The arrival rates are based on the Italian situation, and the pair types are randomly chosen on the basis of the distribution in Tab. 1.

A *matching policy* is a procedure that at each time unit $t > 0$ selects a (possibly empty) set of matching pairs from the pool. Once a pair is matched at time t by a matching mechanism, it leaves the pool and its patient receives the assigned kidney. The matching policies we consider belong to a very special class. More precisely we suppose that the central authority decides to organize optimal matchings at regular intervals of time. The frequency of the exchanges remains fixed and is chosen in advance, and the matching policy organizing the optimal matching every τ units of time is denoted by π_τ. We experimentally compare exchange policies π_τ for different values of τ using the simulator we have developed in C++ using the LPsolve library (http://lpsolve.sourceforge.net) for the solution of the optimization problem every τ units of time. The comparison is made in terms of the number of matched pairs, the average waiting time and other aspects we now describe.

In order to get results easy to interpret from a practical viewpoint, we fix the time unit to be one month, and we choose the other parameters consequently. The arrival rate is of 3 pairs per time unit, which corresponds to a realistic scenario for the Italian situation, for instance. We ran the simulations adopting the policies π_τ for $\tau = 1, 2, 3, 4, 6, 12$. Each experiment is replicated 20 times (for different arrivals) to get more robust results. Confidence intervals are computed for each selected index, setting the confidence level to 95% and achieving an accuracy of at least 0.37. The time horizon is 240 time units (20 years), to which we added an initial tran-

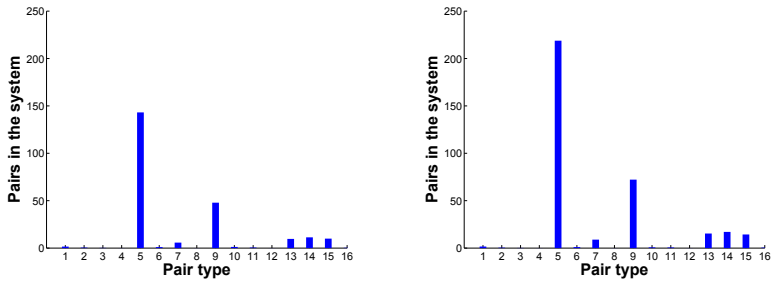


Fig. 1 Average number of pairs in the system for each type, obtained using policy π_6. Results for $TP = 36$ (left) and $TP = 120$ (right).

sient period. We start to evaluate the quantities we are interested in after the initial transient period. We analyze the results of two groups of experiments: in the first group the transient period (TP) is 36 months, and in the second one is 120 months. This choice is due to the fact that this system does not have a steady state since the *underdemanded pairs* form an increasing queue in the system (see [11] and [9]). Therefore choosing the length of the transient period has the effect of determining

a different distribution on the initial state. We are interested in showing the dependence of the performances of the same policy under different initial conditions. The accumulation effect appears, as can be seen clearly in Fig. 1, no matter which policy is adopted. If the transient period is longer, the average number of pairs in the system increases and goes to infinity. This problem is a key point that must be addressed in practice and has also ethical implications. A possible solution would be not to admit to such programs a number of pairs of underdemanded types above a given threshold, since in any case the transplant would be unachievable for them.

It turns out that the average number of pairs in the system per time unit remains practically constant for underdemanded pairs (types $5, 6, 9, 13, 14, 15$, see Tab. 1) and it increases only slightly for the remaining types if τ varies. For instance the average number of pairs of type 1 passes from 0.5 in the case matchings are organized every month to 2.8 in the case they are organized only once a year.

Now, we discuss how the parameter τ affects the percentage of performed transplants using the policy π_τ.

Table 2 Average number of transplanted patients (left). Average waiting time for pairs (0,A) and (A,0), types 2 and 5, respectively (right)

τ	1	2	3	4	6	12
TP=36	52.6	52.5	52.8	53.0	53.4	54.6
TP=120	53.0	52.9	53.3	53.5	53.9	55.3

τ		1	2	3	4	6	12
TP=36	Type 2	0	0.4	0.9	1.3	2.3	4.9
	Type 5	110.2	110.1	110.6	111.0	111.9	115.0
TP=120	Type 2	0	0.3	0.7	1.0	1.6	3.6
	Type 5	128.5	128.1	128.4	128.6	129.3	131.3

As expected, this percentage is approximately an increasing function of τ, varying not too much w.r.t. the considered parameters (Tab. 2 (left)). The fact that this percentage is almost independent of the duration of the transient period ensures that this kind of policies does not suffer from the point of view of the accumulation effect due to the underdemanded pairs. If we look only at the percentage of transplanted patients, the most advantageous policy is the one performing the exchanges only once a year. On the other hand, in the evaluation of a given policy, also the average time that a pair spends in the system is relevant. The average waiting time of those receiving a transplant is considerably different from type to type, and the results agree with the ones regarding the average number of pairs in the system reported in Fig. 1. It turns out that the average waiting time for underdemanded pairs is very long, no matter which policy is adopted, while the average waiting time for the remaining pairs slightly increases as τ grows. As a paradigmatic example we consider pairs of type 2 and 5.

Furthermore, as can be seen from Tab. 2 (right), the situation of underdemanded pairs further deteriorates if the transient period is longer, due to the accumulation effect, as it has been observed in [10], where the results of a 3-year simulation of policy π_1 have been discussed.

Comparing our results with the ones obtained in [1] is not straightforward, since in that paper the average waiting time is not considered, and different policies are compared only on the base of the total number of performed exchanges. In our case maximizing this quantity would imply the use of a policy organizing the exchanges on the last month of the simulation, which is not a viable solution from a practical point of view.

4 Conclusion

In this paper we discussed the results of realistic simulations of several matching policies obtained thanks to a tool, available upon request to the authors. We highlighted that such simulations suggest that KEPs must address *in primis* the problem of underdemanded patients accumulating in the pool without receiving a transplant. Moreover we showed how the timing between a matching and the subsequent one affects the number of performed transplants and the average waiting time w.r.t. pairs characteristics.

The present model can be modified in order to deal with more general situations in which there are more pair types, based also on other genetic or physical characteristics; using a compatibility matrix whose entries can assume real non-negative values enables to model policies that favor the exchanges involving underdemanded pairs, thanks to a (very) small increasing of their *revenues*. Exploiting the flexibility of the tool, further improvements may be in the direction of allowing exchanges involving more than two pairs, via a cyclic series of transplants.

Acknowledgements This research has been partially supported by the Regione Basilicata project "Analisi strategica, modellizzazione matematica ed algoritmi per un programma di scambi di organi" and it has been carried out in collaboration with "Centro Nazionale Trapianti".

References

1. P. Awasthi and T. Sandholm. Online stochastic optimization in the large: Application to kidney exchange, 2009. preprint.
2. K. Cechlàrovà, T. Fleiner, and D. Manlove. The kidney exchange game. In *SOR'05*, pages 77–83, Slovenia, 2005. L. Zadnik-Stirn and S. Drobne, eds.
3. F.L. Delmonico, P.E. Morrissey, G.S. Lipkowitz, J.S. Stoff, J. Himmelfarb, W. Harmon, M. Pavlakis, H. Mah, J. Goguen, R. Luskin, E. Milford, G. Basadonna, M. Chobanian, B. Bouthot, M. Lorber, and R.J. Rohrer. Donor kidney exchanges. *Am J Transplant*, 4: 1628–1634, 2004.
4. K. M. Keizer, M. de Klerk, B.J.J.M. Haase-Kromwijk, and W. Weimar. The dutch algorithm for allocation in living donor kidney exchange. *Transplantation Proceedings*, 37: 589–591, 2005.
5. A. Nicolò and C. Rodriguez-Álvarez. Feasibility constraints and protective behavior in efficient kidney exchange. *FEEM Working Paper*, 31, 2009.
6. C. Petrini, S. Venettoni, and A. Nanni Costa. Il trapianto crossover: aspetti generali e di etica. *MEDIC Metodologia didattica e innovazione clinica*, 15(1): 72–83, 2007.
7. F. T. Rapaport. The case for a living emotionally related international kidney donor exchange registry. *Transplantation Proc.*, 18: 5–9, 1986.
8. A. E. Roth, T. Sönmez, and U. M. Ünver. Kidney exchange. *The Quarterly Journal of Economics*, 119(2): 457–488, 2004.
9. A. E. Roth, T. Sönmez, and U. M. Ünver. Pairwise kidney exchange. *Journal of Economic Theory*, 125(2): 151–188, 2005.
10. D. L. Segev, S. E. Gentry, J. K. Melancon, and R. A. Montgomery. Characterization of waiting times in a simulation of kidney paired donation. *American Journal of Transplantation*, 5: 2448–2455, 2005.
11. U. Ünver. Dynamic kidney exchange. *Review of Economic studies*, 77(1): 372–414, 2009.
12. S. A. Zenios. Optimal control of a paired-kidney exchange program. *Manage. Sci.*, 48(3): 328–342, 2002.

Production Planning for Pharmaceutical Companies Under Non-Compliance Risk

Marco Laumanns, Eleni Pratsini, Steven Prestwich, and Catalin-Stefan Tiseanu

Abstract This paper addresses a production planning setting for pharmaceutical companies under the risk of failing quality inspections that are undertaken by the regulatory authorities to ensure good manufacturing practices. A staged decision model is proposed where the regulatory authority is considered an adversary with limited inspection budget, and the chosen inspections themselves have uncertain outcomes. Compact formulations for the expected revenue and the worst-case revenue as risk measures are given as well as a proof that the simplest version of the production planning problem under uncertainty is NP-complete. Some computational results are given to demonstrate the performance of the different formulations.

1 Introduction

Pharmaceutical companies must obey strict regulations for manufacturing their products. These regulations are usually referred to as Good Manufacturing Practices (GMPs) and are enforced by the national regulatory agencies, who perform inspections to ensure that products are produced safely and correctly. It is therefore important for a company to quantify and manage its resulting risk exposure [3, 2].

Marco Laumanns
IBM Research – Zurich, 8803 Rueschlikon, Switzerland e-mail: mlm@zurich.ibm.com

Steven Prestwich
Cork Constraint Computation Centre, University College, Cork, Ireland e-mail: s.prestwich@cs.ucc.ie

Eleni Pratsini
IBM Research – Zurich, 8803 Rueschlikon, Switzerland, e-mail: pra@zurich.ibm.com

Catalin-Stefan Tiseanu
University of Bucharest, Faculty of Mathematics and Informatics, Bucharest, Romania, e-mail: ctiseanu@gmail.com

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 545
DOI 10.1007/978-3-642-20009-0_86,

Of particular importance for production planning is the risk transfer between products. If only one product fails the inspection, all drugs produced at a site might be considered adulterated. In this paper, we show how these interdependencies regarding non-compliance risk can be formalized and integrated into production planning models based on integer programming.

To derive a suitable model for production planning under uncertainty, we separate the random events faced by the company into two phases:

1. the selection of production sites that are inspected by the regulatory agency, and
2. the subsequent success or failure of each inspection and its effect on the revenue of the products produced at the inspected site.

The actual production decisions have to be taken under uncertainty, before both phases of random events take place.

As the agencies' inspection strategy is typically unknown, we consider the worst case and model the agency as a perfect-information adversary that attempts to minimize the company's revenue. Although this is certainly not the real objective of the regulator, taking a maximin-approach from the point of view of the company is reasonable in such game-theoretic setup where the company has to commit to a plan without knowing the inspection strategy and wants to secure a certain revenue.

For the second phase, we also assume that the outcome of an inspection is uncertain, but the company might have some probabilistic information in this case. In particular, the company is of course aware of its internal production processes and the factors (like water quality, skills and training of workers, results from internal quality control or past inspections) that influence product quality and hence the likelihood to fail an inspection. We thus model the outcome of a given inspection as a Bernoulli-distributed random variable whose parameter (the probability of failing an inspection) is a function of the production site and all products produced at that site. The products produced at a certain site can only generate revenue if the site passes inspection or is not inspected at all. In case of a failed inspection, the site revenue is zero. This is to model GMP rules that products of a site that failed inspection cannot be sold.

The resulting total revenue after inspections consequently becomes a random variable as well. In the sequel, we discuss different risk measures for the total revenue as the objective for production planning and derive corresponding integer programming formulations. In addition, some numerical test are performed on randomly generated instances to compare the performance of the different models when solved with a commercial solver.

2 Worst-Case Revenue under Limited Inspection Budget

Let a company be given with S production sites $s \in \mathcal{S} = \{1,2,\ldots,S\}$. There are P products $p \in \mathcal{P} = \{1,2,\ldots,P\}$, which it has to distribute over the sites in order to optimize a certain risk measure of the total revenue. Each product p yields some

revenue $r_p \in \mathbb{N}$, independently of the site where it is produced. Additionally, each product p carries a hazard $h_{s,p} \in [0,1]$, which is dependent on the particular site s. The regulatory agency, seen as an adversary, can inspect at most K sites. In case of inspection of a site, the probability of it failing the inspection is equal to the maximum over the risks of all products made at that site, which models product adulteration.

We start by considering the worst-case revenue. In the worst case, the individual hazard values do not play a role since the company is assumed to fail each inspection. Therefore, the problem becomes how to distribute products over the sites as evenly as possible, that is: given S sites, and P products, each with a certain revenue r_p, how to allocate them to sites in order to maximize the total revenue when a perfect-information adversary will eliminate the revenues in the K sites with highest site revenue. Let us call the problem to decide whether a certain worst-case revenue target can be achieved, WORSTCASE-COMPANY.

Proposition 1 *WORSTCASE-COMPANY is NP-complete.*

Proof. We use reduction from HALF-PARTITION, which is the problem of deciding whether a given multiset of N numbers can be partitioned into 2 multisets of equal sum, and which is NP-complete. Let $X = \{a_1, \ldots, a_N\}$ be an instance of HALF-PARTITION. Let $\sigma = \sum_{a \in X} a$ and assume w.l.o.g. that σ is even. We can easily transform X into an instance of WORSTCASE-COMPANY by setting $S = 2$, $P = |X|$, $r_i = a_i$ for each product i and $K = 1$. It can now easily be seen that there exists a production plan for instance Y of WORSTCASE-COMPANY of revenue $\sigma/2$ if and only if the X can be partitioned in 2 sets of equal sum.

The following, mixed-integer linear programming formulation for the optimization version of WORSTCASE-COMPANY is a direct formulation, using $x_{p,s}$ as binary decision variables for the allocation of product p to site s, w_s as auxiliary variables indicating inspection of site s, and v_s for the revenue of site s:

$$
\begin{aligned}
\text{Max} \quad & \sum_{s \in \mathcal{S}} v_s & & \\
\text{s.t.} \quad & v_s \leq \sum_{p \in \mathcal{P}} r_p x_{p,s} & & \forall s \in \mathcal{S} \\
& v_s \leq M(1 - w_s) & & \forall s \in \mathcal{S} \\
& \sum_{p \in \mathcal{P}} r_p x_{p,k} + M(1 - w_k) \geq \sum_{p \in \mathcal{P}} r_p x_{p,l} - M w_l & & \forall k \in \mathcal{S}, \forall l \in \mathcal{S} \\
& \sum_{s \in \mathcal{S}} x_{p,s} \leq 1 & & \forall p \in \mathcal{P} \\
& \sum_{s \in \mathcal{S}} w_s = K & & \\
& x_{p,s} \in \{0,1\} & & \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \\
& w_s \in \{0,1\} & & \forall s \in \mathcal{S} \\
& v_s \geq 0 & & \forall s \in \mathcal{S}
\end{aligned}
$$

Letting $M = \sum_{p \in \mathcal{P}} r_p$, the second constraint ensures that the revenue of an inspected site is zero while the third constraint ensures that the inspected sites are those with largest revenue. Further constraints to limit the production capacity of sites or to disallow production of a product at a certain site can be added if necessary.

This formulation has the disadvantage that it introduces additional binary variables for the inspections. The fact that the adversary's problem of choosing the inspections (after the production plan is fixed) can be solved greedily suggests that it could be avoided. To achieve this, we start with a formulation as a bi-level problem that mimics the original staged structure of first the production decision, followed by the inspections:

$$\begin{array}{ll} \max\limits_{x_{p,s}} & \min\limits_{w_s} \sum\limits_{s \in \mathcal{S}} (1 - w_s) \sum\limits_{p \in \mathcal{P}} r_p x_{p,s} \\ x_{p,s} \in \{0,1\} \ \ \forall s \in \mathcal{S}, p \in \mathcal{P} & \sum\limits_{s \in \mathcal{S}} w_s \leq K \\ \sum\limits_{s \in \mathcal{S}} x_{p,s} \leq 1 \ \forall p \in \mathcal{P} & w_s \in \{0,1\} \quad \forall s \in \mathcal{S} \end{array}$$

Since the LP relaxation of the inner minimization problem (the decision on the inspections) is naturally integer, we can work with its dual to form a max-max problem, which can then be formulated as

$$\begin{array}{lll} \max & \sum\limits_{s \in \mathcal{S}} \sum\limits_{p \in \mathcal{P}} r_p x_{p,s} - Kz - \sum\limits_{s \in \mathcal{S}} y_s & \\ \text{s.t.} & z + y_s \geq \sum\limits_{p \in \mathcal{P}} r_p x_{p,s} & \forall s \in \mathcal{S} \\ & \sum\limits_{s \in \mathcal{S}} x_{p,s} \leq 1 & \forall p \in \mathcal{P} \\ & x_{p,s} \in \{0,1\} & \forall s \in \mathcal{S}, \forall p \in \mathcal{P} \\ & z \geq 0 & \\ & y_s \geq 0 & \forall s \in \mathcal{S} \end{array}$$

This is similar to the robust optimization formulation with budgeted uncertainty proposed by [1] and does not need additional binary variables. We call it 'protected' formulation since the dual variables y_s can be seen as protection costs for the site revenues. Additionally, there exists a 'full adversary' formulation, in which all possible responses of the adversary (combinations of inspected sites) are enumerated beforehand and implemented as constraints in the model.

3 Expected Revenue, CVaR, and Product Adulteration

For using the expected value as a risk measure, we can proceed as before. We now have to consider the failure probabilities and thus add auxiliary variables $f_{p,s}$ for

the site hazard (the maximum hazard of the products produced at site s, given that p is produced at s):

$$
\begin{aligned}
\max \quad & \sum_{s\in\mathcal{S}} v_s \\
\text{s.t.} \quad & v_s \le \sum_{p\in\mathcal{P}} r_p x_{p,s} && \forall s\in\mathcal{S} \\
& M(w_s-1)+v_s \le \sum_{p\in\mathcal{P}} r_p(1-f_{p,s}) && \forall s\in\mathcal{S} \\
& \sum_{p\in\mathcal{P}} r_p f_{p,k} + M(1-w_k) \ge \sum_{p\in\mathcal{P}} r_p f_{p,l} - M w_l && \forall k\in\mathcal{S}, \forall l\in\mathcal{S} \\
& (1-x_{p,s})+f_{p,s} \ge h_{q,s} x_{q,s} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P}, \forall q\in\mathcal{P} \\
& \sum_{s\in\mathcal{S}} x_{p,s} \le 1 && \forall p\in\mathcal{P} \\
& 0 \le f_{p,s} \le x_{p,s} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P} \\
& x_{p,s}\in\{0,1\} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P} \\
& w_s\in\{0,1\} && \forall s\in\mathcal{S} \\
& v_s \ge 0 && \forall s\in\mathcal{S}
\end{aligned}
$$

Again, we can dualize the corresponding bi-level formulation to arrive at a compact version without additional binary variables for the inspections:

$$
\begin{aligned}
\max \quad & \sum_{s\in\mathcal{S}}\sum_{p\in\mathcal{P}} r_p x_{p,s} - Kz - \sum_{s\in\mathcal{S}} y_s \\
\text{s.t.} \quad & z + y_s \ge \sum_{p\in\mathcal{P}} r_p f_{p,s} && \forall s\in\mathcal{S} \\
& (1-x_{p,s})+f_{p,s} \ge h_{q,s} x_{q,s} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P}, \forall q\in\mathcal{P} \\
& 0 \le f_{p,s} \le x_{p,s} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P} \\
& \sum_{s\in\mathcal{S}} x_{p,s} \le 1 && \forall p\in\mathcal{P} \\
& x_{p,s}\in\{0,1\} && \forall s\in\mathcal{S}, \forall p\in\mathcal{P} \\
& z \ge 0 \\
& y_s \ge 0 && \forall s\in\mathcal{S}
\end{aligned}
$$

In contrast with the other two risk measures, which allowed for a simple polynomial greedy algorithm for the adversary, once the production was known, other risk measures such as the conditional value-at-risk (CVaR) are far more difficult to handle. The reason is that one can no longer look at the individual site failures independently, but has to consider the exact distribution over the 2^K possible different outcomes. In fact, it can be proven that only computing the CVaR for a given production plan and inspections is NP-hard, i.e., exponential in the number of inspected sites. Therefore we refrain from giving a corresponding MIP formulation as it would be of exponential size and not very useful in practice.

4 Computational Results

For our experimental results, we generated random instances of different size with product revenues uniformly distributed in the range $[0, 200]$ and product risks uniformly in the range $[0, 0.1]$. For every problem size (test case), the results in Table 1 give the CPU time to solve the problem to optimality, averaged over 10 random instances. The following results were obtained using an Intel T2600 processor at 2.16 GHz and 2 GB of RAM, using the MIP solver of CPLEX version 12.

Table 1 Solution times for the different models. WC stands for the worst-case risk measure and E for the expected value. As second identifier, D stands for direct formulation, F for full adversary, and P for the protected formulation. The dash (-) means that at least one instance could not be solved within the given time limit of one minute.

Case	S	P	K	WC-D	WC-F	WC-P	E-D	E-F	E-P
1	6	12	3	6.62	0.54	3.72	3.37	0.25	0.24
2	6	12	6	0.01	0.01	0.01	0.12	0.09	0.11
3	7	14	3	-	7.54	-	20.91	0.93	1.04
4	7	14	7	0.01	0.01	0.01	0.20	0.13	0.19
5	8	16	3	-	-	-	-	4.52	3.46
6	8	16	4	-	-	-	-	1.97	1.44
7	9	18	3	-	-	-	-	8.84	4.52
8	9	18	5	-	-	-	-	8.93	3.69
9	10	20	3	-	-	-	-	14.85	6.54
10	10	20	5	-	-	-	-	18.08	5.04
11	12	24	3	-	-	-	-	-	-
12	12	24	6	-	-	-	-	-	37.95

5 Conclusions

We have studied production planning for pharmaceutical companies under the risk of failing inspections that are undertaken by the regulatory authorities to ensure good manufacturing practices. A staged decision model was proposed where the authority is considered an adversary with limited inspection budget, and that the chosen inspections themselves have uncertain outcomes. We have given compact formulations for the expected revenue and the worst-case revenue as risk measures and have indicated why it is hard to find a reasonable MIP formulation for the CVaR.

References

1. D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52: 35–53, 2004.
2. A. Elisseeff, J. P. Pellet, and E. Pratsini. Causal networks for risk and compliance: Methodology and application. *IBM Journal of Research and Development*, 54(3): 6:1–6:12, 2010.
3. E. Pratsini and D. Dean. Regulatory compliance of pharmaceutical supply chains. *ERCIM News*, 60: 51–52, 2005.

A Model for Telestroke Network Evaluation

Anna Storm, Stephan Theiss, and Franziska Günzel

Abstract Different telestroke network concepts have been implemented worldwide to enable fast and efficient treatment of stroke patients in underserved rural areas. Networks could demonstrate the improvement in clinical outcome, but have so far excluded a cost-effectiveness analysis. With health economic analysis lacking, current telestroke reimbursement by third-party payers is limited to special contracts and not included in the regular billing system. Based on a systematic literature review and expert interviews with health care economists, third-party payers and neurologists, a Markov model was developed from the third-party payer perspective. In principle, it enables telestroke networks to conduct cost-effectiveness studies, because the majority of the required data can be extracted from health insurance companies' databases and the telestroke network itself. The model presents a basis for calculating the telestroke value creation potential and for cost sharing approaches among different third-party payers.

1 Introduction

In western industrialized countries, stroke is the third leading cause of death as well as the main reason for significant long-term disability and high health care treatment costs. According to current guidelines, specialized approved stroke units provide the best treatment for stroke patients, but due to severe lack of neurologists they cannot be established especially in rural areas. Therefore, in many countries, different telemedical network concepts have been implemented to improve acute

Anna Storm, Franziska Günzel
Faculty of Economics and Management, Otto von Guericke University Magdeburg, e-mail: [Anna.Storm | Franziska.Guenzel]@ovgu.de

Stephan Theiss
Faculty of Electrical Engineering and Information Technology, Otto von Guericke University Magdeburg, e-mail: Theiss@uni-duesseldorf.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_87,

stroke care in regions where no stroke units are available [4]. In these "telestroke" networks, expert neurologists based e.g. in tertiary care stroke centers provide 24/7 teleneurological consultation services to remote small primary care hospitals. Using a high-quality two-way audio and video transmission link and dicom interface, these expert consultants can review the patients' brain scans (CT or MRI) and remotely examine their neurological status. Together with the on-site emergency physician in the primary care hospital, they decide on a treatment plan for the patient—in particular about the most effective acute ischemic stroke therapy, administering the thrombolytic clot-busting drug t-PA. Furthermore, the experts select patients eligible for special neurosurgical or neuroradiological intervention and assist in the transfer to the respective specialty centers.

Several clinical studies by telestroke networks have demonstrated significant improvement of telestroke patient management and outcome, but did not focus on health economic aspects. However, third-party payers ask for cost-effectiveness studies of telestroke networks prior to including telemedical treatment into the regular billing system.

The aim of the present study was to create a model for the joint evaluation of both economic and medical aspects in telestroke. The developed model builds on data that is in principle available for all networks. The results of such an evaluation should demonstrate telestroke value creation potential to third-party payers and thus provide a basis for the calculation of an economic based price for telestroke care.

2 Telestroke Model Design

Methodology This study was divided into two successive steps. First, a systematic MedLine and PubMed literature review was carried out to determine the methodological approaches that were used to evaluate (tele-) stroke treatment. In a second step, actual models were developed and subsequently discussed in open-ended expert interviews regarding the degree of modeling detail, states and outcome measures. To this end, four experts with a background in (tele-) stroke care and management were selected in the areas of health care economics, insurance and neurology.

Markov Model Approach for Telestroke Health Economics In the field of stroke health care evaluation, several different modeling techniques are applied—especially (semi-) Markov models but also decision trees and spreadsheet models. To select an appropriate approach, the attributes of the existing evaluation methods were revised and compared with telestroke characteristics. Using decision modeling, it is possible to follow the course of a disease, combine different data sources, and extrapolate primary data for a longer time horizon, as it is required for stroke. In the realm of decision modeling Markov models can describe time-dependent properties as the development of a disease and the incidence of recurring events for fixed, e.g. annual, periods and are thus well suited for telestroke evaluation. Markov models can either consider average patients sharing the same characteristics in a cohort

model or a random sample of individual patients allowing for higher diversification between patients in micro simulations [5]. Stroke epidemiology studies show that stroke patients form a relatively homogeneous elderly population sharing similar risk factor profiles. Although in principle transition rates between the states of a Markov model may depend on individual patient characteristics, consistent data is rare and this level of detail not required, so that telestroke evaluation can be modeled by cohorts.

Separate Study Cohorts for Stroke Types "Ischemia" and "Hemorrhage" To properly determine the cost-effectiveness of telestroke care all patient groups treated need to be taken into account. Since current telestroke networks aim at improving outcome both for patients with acute ischemic and hemorrhagic stroke, a Markov model for telestroke networks must include both stroke types. In particular, the costs incurred and the transition rates to patient outcome states vary widely between the two stroke types [3]. This calls for using two separate study cohorts for ischemic and hemorrhagic stroke patients on the top level of the Markov model. Besides, the transition rates to different outcome states for ischemic stroke patients strongly depend on the state-of-the-art t-PA therapy, entailing a third level splitting of the ischemic stroke cohort in network hospitals into patients receiving t-PA treatment or not. The rate of t-PA treatment on general wards is extremely low, so that no t-PA arm is included in this branch of the Markov model.

Telestroke evaluation relative to standard "baseline" patient care In a telestroke cost-effectiveness study, telestroke network concepts need to be compared to the current standard stroke patient care as reference process. Since telestroke networks do not aim to compete with specialized stroke units but rather complement these in underserved areas, this reference process is represented by the stroke patient care provided on a general ward in a primary care hospital without resident neurologist or other stroke expert. Moreover, both primary care hospitals and expert centers establishing and running telestroke networks incur personnel costs for consultant neurologists and additional nursing staff as well as technology, IT network and maintainance costs. This implies a two-armed Markov model for each stroke subtype: the initial stroke patient cohorts are randomly split into two equally large cohorts of patients admitted to either a general ward ("control arm") or to a telestroke network hospital ("intervention arm"). This choice of Markov model design is in line with the clinical outcome studies performed by the currently largest German telestroke network, TEMPiS.

Care Level: Practical Outcome Parameter for Third-Party Payer Perspective The determination of relevant costs and clinical outcomes to be considered is highly dependent on the chosen perspective. Healthcare economic literature distinguishes four different perspectives–societal, patient, health care provider and third-party payer, and usually recommends the societal perspective for cost-effectiveness analyses [1]. However, experts suggested to use a simple and robust model to present a calculation base to third-party payers–in particular both health care and nursing care insurances reimbursing the major part of total stroke costs–in order to facilitate

the inclusion of telestroke care into the billing system. Therefore, the Markov model estimating telestroke cost saving potential was designed from the third-party payer perspective.

Given the chosen perspective, appropriate outcome measures were selected. Functional status following the index stroke event may predict patients' dependency and has economic impact. The most reliable and frequently used scales are the Barthel Index (BI) and modified Rankin Scale (mRS), assessing handicap in the activities of daily living [2]. Experts agreed that BI and mRS describe valid and appropriate outcome measures, but cautioned that there is no consensus about which scale should be used. Especially in primary care hospitals various scales—or none at all—are adopted. Therefore, as an alternative relevant and available measure the nursing insurance care level was selected, an integer scale quantifying a patient's disability between 1 (≥ 1.5 hours care per day) and 3 (≥ 5 hours care per day) that is calculated from a combination of BI and the early rehabilitation index in the German health care system. Additional outcome states healthy (without care level) and stroke related death were integrated. Using the care level as outcome quantity has medical and economic benefits: it correlates with the valid measure BI, determines the main costs incurred by nursing care insurances and is easily available from the health insurance database.

Definition of the Markov Model and its Data Sources The developed Markov cycle tree is displayed in figure 1. The ischemic or hemorrhagic stroke patient cohorts entering the model are admitted either to a network hospital or a general ward. In the network hospital, the treatment of ischemic patients is differentiated in t-PA or no t-PA treatment. After acute treatment, patients are classified according to their care level or may experience stroke related death. Patients may remain in their states, have a recurrent stroke, be readmitted to the hospital and thus cycle through the model again or experience a non-stroke death. If the stroke etiology is not carefully diagnosed and treated, patients may experience a recurrent, possibly fatal stroke (cumulative risk within 5 years about 25%). The prevention of recurrent stroke is a major focus of telestroke treatment, because it avoids higher treatment costs, outcome worsening and early death, and must therefore be included in the model. Death constitutes the absorbing state and can occur at any point. Age-adjusted non-stroke related death (population mortality) was included in the model for a realistic consideration of time-dependent all case mortality.

Despite the methodological recommendations from health care literature to analyze the full profit of the treatment with a lifetime horizon, experts suggested a time horizon of a maximum of five years, because data for lifetime evaluations are increasingly uncertain and third-party payers are rather interested in a short-term view.

While developing the model, care was taken to ensure the availability of all data items included. To conduct a study the following data sources should be used: health care insurance (e.g. direct cost data), hospital/telemedicine network (e.g. t-PA treatment) and Federal Statistical Office (e.g. age-specific death rates).

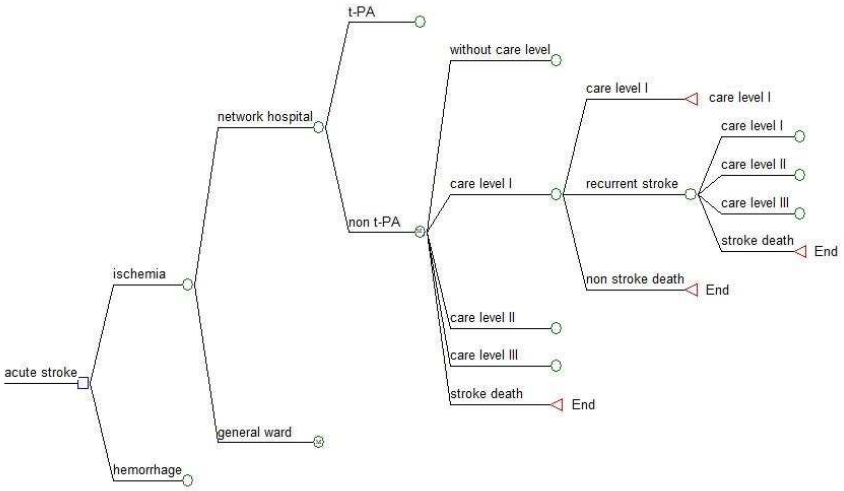


Fig. 1 Markov cycle tree for health economic evaluation of telestroke.

Telestroke Value Creation Potential This paper suggests to calculate different medical and economic outcome measures to evaluate a telemedicine network performance. From the clinical perspective e.g. the number of life years gained, the proportion of patients without severe disability or prevented recurrent strokes could be calculated. A relevant economic outcome parameter is the difference of the overall health care costs between the two treatment alternatives. An incremental cost-effectiveness ratio (ICER) analysis allows the assessment of costs per unit benefit. Since rehabilitation and nursing dominate the stroke treatment costs, introduction of telestroke treatment will likely lead to cost savings for third-party payers in the long run. It is a major feature of the presented approach to permit determining the value creation potential of telestroke care by comparing the treatment and follow-up costs in both telestroke and baseline arm. Furthermore, health insurance and nursing care costs can be extracted separately, offering the basis for cost sharing-approach.

3 Discussion

In this paper, a health economic Markov model for telestroke was developed permitting cost-effectiveness calculation and demonstration of value creation potential. As suggested by Earnshaw et al (2009), the complexity was kept to a minimum without making a large number of assumptions. The majority of the data required for running the model is available from health insurances and telestroke networks, and no additional studies are needed. The model assumes a third-party payer perspective and is thus limited to direct costs. It probably underestimates the effect of telestroke,

since potential clinical improvement for stroke mimics and effects of immediate patient transfer to tertiary care specialty centers have not been included. In this sense, the added value predictions are conservative from the third-party payer perspective. The presented model can in principle also be applied to stroke care in other countries provided similar nursery care level classification and health economic data is available. A more thorough calculation will have to include a sensitivity analysis of the outcome under variation of the input parameters (costs and transition probabilities).

4 Conclusion

Using available data from telestroke network providers and public health insurances, telemedicine networks and third-party payers can jointly apply the presented Markov model framework to evaluate the health economic value of different telestroke network concepts in a given socioeconomic environment. This insight into potential cost-savings should invoke a dialog on cost sharing and telestroke reimbursement strategies between third-party payers. The demonstration of telestroke value creation potential may significantly increase its use for the benefit of stroke patients in underserved areas.

Acknowledgements The project TASC (Telemedical Acute Stroke Care) receives financial support from the German Ministry of Education and Research (BMBF) within the framework of the program "ForMaT - Research for the Market in Teams" (ID 03F01242).

References

1. M.R. Gold, J.E. Siegel, L.B. Russel, and M.C. Weinstein. *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York, Oxford, 1996.
2. C. Haacke, A. Althaus, A. Spottke, U. Siebert, T. Back, and R. Dodel. Long-term outcome after stroke – evaluating health-related quality of life using utility measurements. *Neurology*, 37(1): 193–198, June 2000.
3. S.D. Reed, D.K. Blough, K. Meyer, and J.G. Jarvik. Inpatient costs, length of stay, and mortality for cerebrovascular events in community hospitals. *Neurology*, 57(2): 305–314, July 2001.
4. L.H. Schwamm, R.G. Holloway, P. Amarenco, H.J. Audebert, T. Bakas, N.R. Chumbler, R. Handschu, E.C. Jauch, W.A. Knight, S.R. Levine, M. Mayberg, B.C. Meyer, P.M. Meyers, E. Skalabrin, and L.R. Wechsler. A review of the evidence for the use of telemedicine within stroke systems of care: a scientific statement from the American Heart Association/American Stroke Association. *Stroke*, 40(7): 2616–34, July 2009.
5. F.A. Sonneberg and J.R. Beck. Markov models in medical decision making: a practical guide. *Med Decis Making*, 13(4): 322–38, April 1993.

III.5 Simulation and System Dynamics

Chair: Prof. Dr. Grit Walther (TU Braunschweig)

Decision making in dynamically complex systems requires the application of modern methods for systems analysis and simulation. We therefore invite original contributions, which address the development as well as the application of analysis and simulation of such dynamic systems from a theoretic as well as a practical point of view. Thereby, we encompass methods of discrete event simulation as well as continuous simulation methods.

Topics include exploratory analysis and experimental design, heuristics and algorithmic engineering within complex systems, agent-based modeling, calibration, validation and verification techniques as well as uncertainty within systems and robustness of solutions.

Applications might focus on every kind of dynamically complex systems - (socio-) economic as well as technical and natural systems, and strategic as well as operational planning tasks might be tackled.

Sub-Scalar Parameterization in Multi-Level Simulation

Marko A. Hofmann

Abstract Sub-scalar parameterization refers to substituting processes that are too small-scale or complex to be physically represented in a simulation model by parameters. A typical example for a sub-scalar parameterization is the representation of clouds in climate models. Unfortunately, not all of these parameters can be measured directly. Hence, it is often necessary to calculate sub-scalar parameters (for the primary simulation) using additional models, like special small-scale simulations (secondary simulations). In many applications a dynamic exchange of data between both simulation during runtime is necessary: The high-resolution model iteratively calculates new parameter values using information from the aggregated model. Using such an approach in the climate example, the secondary model would calculate cloud distributions on the basis of simulation results from the primary climate model (e.g. based on global or local average temperatures). However, there is a certain amount of inherent uncertainty in the data flow from aggregated to a high-resolution models. If the output of the secondary high-resolution model is sensitive to this uncertainty, sub-scalar parameterization is at least questionable. The paper formally defines this problem in order to systemize its investigation.

1 Introduction: Verbal Description

Sub-scalar parameterization in simulation models refers to the method of substituting processes that are too small-scale or complex to be physically represented in the model by parameters. This can be contrasted with other processes that are explicitly resolved within the simulation models. Sub-scalar parameterization is indispensable for modeling and simulation since it curtails the complexity of models without completely ignoring the important effects of processes below the level of explicit resolution.

Dr. Marko A. Hofmann
ITIS, Werner-Heisenbergweg 39, D-85577 Neubiberg, e-mail: marko.hofmann@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_88,

A typical example for such a sub-scalar parameterization is the representation of clouds in climate models. Most climate models have a resolution of about 100 km. However, an average cumulus cloud has a scale of less than a kilometer, and would require a grid even finer than this to be represented physically as a dynamic process. Therefore the effects of such clouds are modeled as parameters, representing their albedo, for example. (A critical review of cloud modeling in climate models can be found in [5].)

One of the most important aspects of parameterization is how the substituting parameters are obtained. Some of these parameters can be measured directly (as a current average percentage of cloud covering, for example), while others, in general, defy empirical methods (predictions, for example). This is essentially done using models which can be rather simple like a single equation, or highly sophisticated like complex small-scale simulations. In the latter case, the whole approach is a special kind of multi-level simulation, in which the interface between the simulation models of different resolution is limited to a few parameters. With respect to the parameterization the primary simulation model (representing the macro-phenomena of interest, e.g. climate change) is an aggregated model whereas the model that calculates the parameter is a high-resolution model (called secondary simulation).

In some applications such parameters have to be calculated only once (as initial values), in others they have to be calculated dynamically throughout the simulation time of the aggregated simulation. In the latter case an exchange of data between primary and secondary simulation is necessary: In such applications data produced during a simulation run of the aggregated model has to be transformed or "disaggregated" for the high-resolution model. The high-resolution model iteratively calculates new parameter values using information from the aggregated model. Using such an approach in the climate example, the secondary model would calculate cloud distributions on the basis of simulation results from the primary climate model (e.g. based on global or local average temperatures). Subsequently, the secondary model would provide the primary model with new cloud coverage predictions.

The idea of linking simulation models of different resolution via sub-scalar parameters can be applied recursively. It is a promising method of reducing the complexity of huge target systems. Some scientist even claim that by using such parameter linked simulations of different resolution (for which I suggest the abbreviation: *PLMLS*; *p*arameter *l*inked *m*ulti *l*evel *s*imulation) it might be possible to master the complexity of real world system currently beyond our skills [1].

However, disaggregation and subsequent recalculation of sub-scalar parameters is by no means trivial. Every disaggregation necessarily includes an arbitrary choice, since the higher resolution model incorporates more information than the aggregated model can provide. With other words, there is a certain amount of inherent uncertainty in the data flow from an aggregated to a high-resolution model. If the output of the secondary high-resolution model is highly sensitive to this uncertainty, sub-scalar parameterization is at least questionable, in extreme cases even useless.[1] This problem is called the *s*ensitivity *p*roblem of *d*isaggregation and *a*ggregation (*SPDA*).

[1] High sensitivity signifies that little input variation causes significant output variation.

2 Formal Description of *PLMLS*

The following formal description is reduced to the absolutely essential attributes of *PLMLS* with respect to the *SPDA*. See Fig. 1 to follow the definitions easily.

A ***target system*** T is any kind of real world system in which a phenomenon of interest is located.

A ***primary simulation*** PS is a model of the target system T. It is defined by a set of models states $s_1, \ldots s_n$, a sub-scalar parameter q and a function S_P with is interpreted as a simulation step within PS:[2]

$$S_P(s_1 \ldots s_n, q) \mapsto (s_1', \ldots s_n', q) \tag{1}$$

A ***secondary simulation*** SS is a model of a specific aspect of the target system T. It is defined by a set of models states $r_1, \ldots r_n$, a sub-scalar parameter p and a function S_S with is interpreted as a simulation step within SS:[3]

$$S_S(r_1 \ldots r_m, p) \mapsto (r_1', \ldots r_m', p) \tag{2}$$

The model states $r_1, \ldots r_m$ of SS are interpreted as a refinement (a higher resolution modelling) of an aspect of the target system modelled by q in PS. Additionally, the higher resolution of q in SS may depend on the model states of PS (context information). This process is called ***disaggregation***, if no further information is involved.[4] Consequently, disaggregation D is defined as:

$$D(s_1', \ldots s_n', q) \mapsto (r_1, \ldots r_m) \tag{3}$$

For the first simulation step, the sub-scalar parameter q of PS can be calculated from data from the target system T. For further steps, (indicated by *PS'* and *PS"* in Fig. 1) it has to be calculated using output from the secondary simulation. This process is called ***aggregation***, since high resolution information in SS is reduced for PS. It can be defined by:

$$f_a(r_1', \ldots r_m') \mapsto q' \tag{4}$$

Hence, a ***PLMLS*** is a simulation system in which a primary simulation P_S is linked to a secondary simulation S_S via the parameter q calculated using an aggregation function f_a, which depends on a disaggregation relation D.

The formalism can be illustrated with the climate modeling **example**.

T : global future climate
PS: climate simulation model.

[2] Notice that the sub-scalar parameter q is not affected by the primary simulation, whereas the model states might change during a simulation step.

[3] Notice that the sub-scalar parameter p is only necessary to indicate the possible recursiveness of the problem.

[4] Note that disaggregation is a higher resolution modeling, without actually using additional high resolution data from the target system (since it is unavailable).

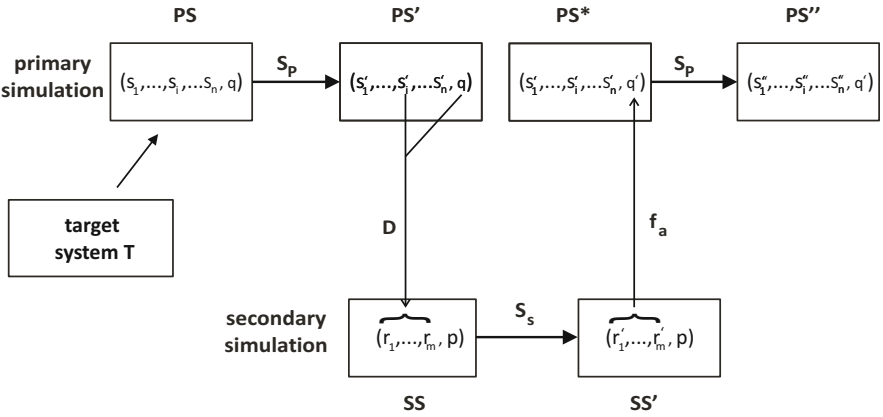


Fig. 1 Parameter linked multi-level simulation

SS: cloud distribution model.
q : percentage of cloud coverage.
s_i : global average humidity (a simplified context).
r_j : cloud coverage in certain region j.
p : local average temperatures.

For the initial simulation step the climate model (PS) can use observational data to quantify q. However, since PS simulates the future, it will be necessary to recalibrate q according to a changed environment (e.g. global average humidity), because these changes affect q. In the example, we assume that this recalibration is done by a cloud distribution model (SS) using the initial value of q and the changed global average humidity as input (which has to be disaggregated into regional cloud distributions $r_1, \ldots r_m$ of SS). The results of SS (new regional cloud distributions) are aggregated in order to provide PS with an updated q' value for the global percentage of cloud coverage, a value which should be more adequate for the projected models states of PS ($(s'_1, \ldots s'_n)$).

3 Formal Description of SPDA

Fig. 1 reveals that there is a gap between two simulation steps S_P in the primary simulation (from *PS'* to *PS**). The change from $(s'_1, \ldots s'_n, q)$ to $(s'_1, \ldots s'_n, q')$ is discontinuous within *PS*, and one might ask how sensitive *PS* is with respect to q. Sensitivity is a measure of how the variation (in the sense of uncertainty) in the output of a model can be apportioned qualitatively to variation in the input of the model [3]. For reasons of simplicity, let us assume that the *PLMLS* consists of only one simulation step, and that the model output $\omega = \phi(s'_1, \ldots s'_n, q')$ is a single scalar

(the global average temperature in 50 years, for example). Thus, we are interested in the variation of ω with respect to the variation of q. Ignoring many details this can be formalized by $\Delta(\omega) = f(\Delta(q))$. Since each application of a simulation model has a ***tolerable threshold of uncertainty with respect to the model purpose***, we can configure $\Delta^*(\omega) := c$ as a such a critical value (a constant, again for reasons of simplicity).[5] Such a threshold is exceeded, if ω is highly sensitive to q:

$$f(\Delta^*(q)) \geq \Delta^*(\omega) = c \tag{5}$$

With other words, there is a critical variation of q with respect to ω. Applied on the climate example, this relation translates to: A certain variation of the percentage of cloud coverage ($\Delta^*(q)$) might result in an variation of the predicted global average temperature in 50 years (ω) which is intolerable for practical decisions.

However, disaggregation and subsequent aggregation always result in uncertainty of the sub-scalar parameter q in *PLMLS*, since the disaggregation relation $D(s_1', \ldots s_n', q) \mapsto (r_1, \ldots r_m)$ is not definite. A global percentage of cloud coverage q and a global average humidity s_i, for example, can be disaggregated into various concrete cloud coverage patterns. These different input patterns ($(r_1, \ldots r_m)$) can easily lead to different output patterns of *SS* ($r_1', \ldots r_m'$), which may be aggregated into different values of q (see Fig. 2, illustrating the problem with the cloud coverage example). Thus, without further information about the appropriate high resolution in *SS*, ***disaggregation and subsequent aggregation in PLMLS cause arbitrary variance*** $\Delta^{Disagg}(q)$ (25 % in Fig. 2). Such a variation might be greater than the tolerable threshold for q (assuming monotony of f):

$$\Delta^{Disagg}(q) \geq \Delta^*(q) \quad \text{resp.} \quad f(\Delta^{Disagg}(q)) \geq \Delta^*(\omega). \tag{6}$$

This effect is not simply theoretical. It is well-known from military simulation models, and has been first described by [4] without coining a special term for the phenomenon. Schaub demonstrated that aggregated combat simulation models (*PS*), which are based on Lanchester equations (differential equations of a special type), and high resolution combat simulation models (*SS*), which are based on single shot models cannot be satisfyingly linked via sub-scalar parametrization of the attrition processes (generation of the Lanchester values (q) from the single shot models). In all approaches he studied the sensitivity of the primary simulation output (force ratio of two opposing forces) was much too sensitive to variations introduced by the disaggregation and further aggregation of the Lanchester values q and the context information about terrain (s_i) and military situation (s_j). Consequently, *PLMLS* based on attrition processes has been completely dismissed. *PLMLS* based on reconnaissance processes, on the contrary, suffers much less from *SPDA*, and is intensively used in military simulation [2]. Instead of using q as a parameter of attrition, this approach defines q as a number of weapon systems located at a specific position. Although, there too, is an information loss from higher resolution models (with many individual locations of vehicles) to the single positioning in the aggregated

[5] In the climate model this could be a variation of more than 2 degrees Celsius, for example.

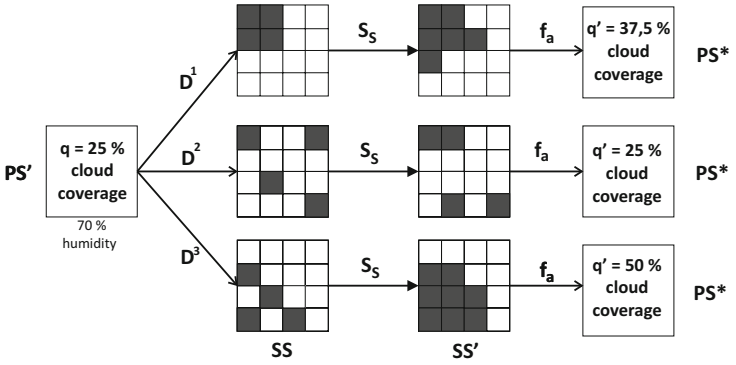


Fig. 2 Ambiguous disaggregation, secondary simulation and subsequent aggregation leading to variance in q'

model, the effect of *SPDA* is much smaller, because unit and terrain representation of the aggregated model can be adjusted to each other. Hence, the degree of *SPDA* in *PLMLS* is not only a matter of domains, but also of choosing an appropriate parameter q. This might be an instructive (negative and positive) example for many similar approaches in various fields of decision supporting simulation.

4 Conclusion

The discussion and formal framework introduced in this paper might help to

- first, raise awareness of the problem of disaggregation and subsequent aggregation in parameter linked multi level simulation,
- second, systemize the problem's investigation, and
- third, help to avoid a serious systematic error inherent in this technique.

References

1. G. Gramelsberger. *Computerexperimente. Zum Wandel der Wissenschaft im Zeitalter des Computers*. Transcript, Bielefeld, 2010.
2. B. Probst, S. Grossmann, M. Hofmann, and M. Schuler. *Anbindung verteilte Simulation an Führungsinformationssysteme*. ESG, München, 2009.
3. A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, S. Saisana, and S. S. Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, New Jersey, 2008.
4. T. Schaub. *Zur Aggregation heterogener Abnutzungsprozesse in Gefechtssimulationsmodellen*. Universtät der Bundeswehr Neubiberg, Neubiberg, 1991.
5. G. L. Stephens. Cloud feedbacks in the climate system: A critical review. *Journal of Climate*, 18: 237–273, 2005.

Exploring Anti-Counterfeiting Strategies: A Preliminary System Dynamics Framework for a Quantitative Counterstrategy Evaluation

Oliver Kleine and Marcus Schröter

Abstract Today, product counterfeiting and piracy are fully recognised as essential business risks in nearly any industry domain. Remedies to end this threat have been researched more than 30 years – yet, this scourge is prevailing. Recent research indicates that decision makers have not yet been enabled to fully understand the scope and implications of the phenomenon in a strategic management context. In particular decision support tools stand for an immediate research and management need. This short contribution aims at closing this gap by proposing a System Dynamics framework for counterstrategy evaluation.

1 Introduction

Product piracy and related phenomena are still rampaging global economies - reported financial losses still enormous. While industry bodies and governments fear the overall impacts on their economies and welfare, it is in particular the recent increase in scope and professionalism of pirate activities that have given private businesses reasons for increased concern. Industry domains that were once thought immune of product piracy, such as the German medium and high-tech industrial goods industry, are now frequently targeted [4]. Simultaneously, the "line" between licit and illicit activities has blurred and therefore blunted counterstrategies solely relying on legal remedies. Furthermore, the planning and deployment of counterstrategies usually draws significantly on financial, personnel and technological resources – binding them on a long term basis and may even exceed the potential

Oliver Kleine
Fraunhofer ISI, Breslauer Str. 48, 76139 Karlsruhe e-mail: oliver kleine@isi.fraunhofer.de

Marcus Schröter
Fraunhofer ISI, Breslauer Str. 48, 76139 Karlsruhe e-mail: marcus.schroeter@isi.fraunhofer.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_89,

financial losses. The risks at stake call for a systematic decision support in this domain.

Today's research is full of good advice when it comes to decide on "what" to do and "how" to do it. Yet, product piracy is prevailing. It is the fundamental question of "whether" and "when" to act that is still unsolved – even after more than 30 years of research. In fact, recent research indicates that is in particular the lack of research in determining the best strategy in terms of conditions, implications, limitations and cost-benefit ratio that stand for an immediate management need, and underlines that "the current state of our understanding ... is, on balance, poor" [1]. Strategic decision support has not yet gone beyond management guidelines and process models – analytical means to analyse the impact of product piracy on business a priori do not exist [3].

Thus, in this short article we want to advance research in this domain by making the case for System Dynamics (SD). With that we not only seek to overcome the deficits in instrumental decision support but specifically aim at providing new means to analyse the impact of product piracy and potential counterstrategies – not as a substitute but as a complement to empirical research designs.

2 On the Impact of Product Piracy

The vast majority of scholars argues that the main problem in dealing with product piracy and counterfeiting is managerial insight. It appears as if the "traditional" mental models of strategic decision makers on competition are flawed when it comes to product piracy. However, a thorough understanding of the fundamental mechanisms is a prerequisite for compiling countermeasures into a conclusive counterstrategy [2]. Therefore, management must acknowledge that competition in a piracy environment is in many ways the same as it used to be – but has changed fundamentally in others. According to current research this concerns in particular three problem domains:

- *The effects of piracy on business are all but clear*. The negative effects such as loss in revenues, brand value or even economies of scale may be (partially) set off by positive impacts as bandwagon, network and (technology) lock-in effects as well as the establishment of market barriers to regular competitors.
- *The basic market mechanics remain unchanged*. Thus, pirates must be considered as a new type of competitor, who deliberately violate legal boundaries and can therefore resort to other strategic options in competition than regular competitors can do. These are basically derived from the deliberate and unauthorised use of intellectual property rights (IPR) - the extent of the IPR violation determines whether or not they can operate openly in the market.
- *Product piracy is as much a demand side problem as it is one for the supply side*. It would be naive to assume that all pirate purchases actually happen unwittingly and unwillingly. On the contrary, research indicates a significant share of customers who are making their buying decisions in this respect wilfully. This

implies further that they are able to readily compare the net utility of the pirate's product – not only to the original but also to the products of other competitors.

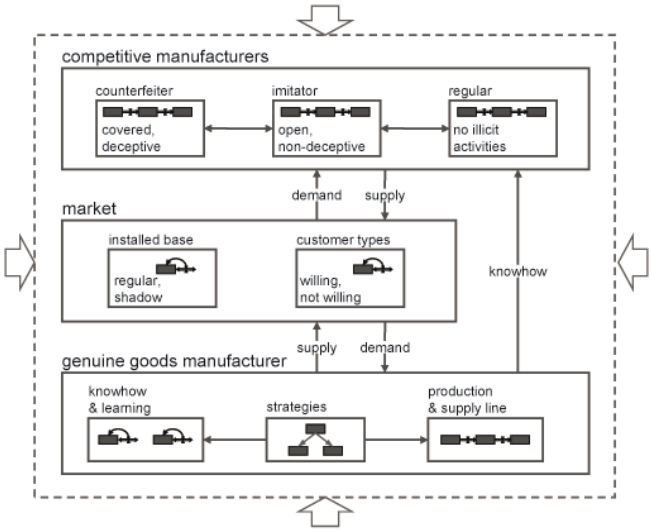


Fig. 1 The "piracy game": a mental model on competition in a piracy environment

Figure 1 summarises the above in a first mental model – the "piracy game" – and includes what we find are the characteristic features of a competition in a piracy environment: These are in particular different types of competitors who differ in how they conduct their activities, customers who either buy pirate products willingly or unwillingly, and the existence of an installed base of pirate products.

To conclude, product piracy will inevitably impact competition – and because of the features mentioned above, probably in other ways than expected. Pirate competitors and customers actively seeking their products have introduced "new" cause and effect relationships that have not yet been recognised in traditional mental models on competition, making an intuitive assessment all but impossible. However, it is not so much their difference in impact relevance and inter-system coupling that is responsible for counter-intuitivity in a piracy competition than it is their impact delays. It is the overall impact of these relationships on a tightly coupled socio-technical system that has to be analysed here. We find that SD might proof a practicable instrument to do so. Its application to answer questions related to strategic production management problems has long since been proved. As an instrument it not only helps to answer a specific problem but explicitly fosters second order learning processes in order to change decision maker's mental models. A SD model deployed in managerial context as a "simulator" could provide means to initiate the necessary learning processes – leading to a better problem insight, and thus, to better decisions.

3 A System Dynamics Framework

In terms of a SD modelling process, Figure 1 also serves as a subsystem diagram defining the problem and model boundaries. The next step would be its formal description. For this purpose we note the following:

- The analytical focus of problems related to the management of product piracy is more strategic than operational, investigating processes at a company and market rather than at a sub-firm level.
- The cause and effect relationships either relate to changes in physical material flows such as (produced and sold) products or to informational flows such as perceptions. The delays in these changes are rather measured in weeks and months than in hours or days. Thus, even the "short-termed" effects in a piracy competition are rather "long-termed" compared to processes on an operational production level, allowing us to approximate discrete material flows continuously.
- Since it is a second order learning process that we must achieve in order to change mental models on a piracy competition, a meaningful qualitative behaviour of the system is more important than the accurate reproduction of empirical data. This is in particular relevant as reliable, empirical data is not available yet.

Thus, we not only find SD a fitting modelling approach but in particular that any of the relevant cause and effect relationships can be sufficiently approximated by rather simple N^{th} -order linear differential equations such as $\frac{\partial S_i}{\partial t} = \sum_i^N \alpha_i S_i$, where S_i denotes the state variables of the system.

Based on that and on Figure 1 the model formulation was straight forward by implementing differential equations such as in (1) to describe the behaviour for each relevant model element. It is structured into three main sectors: A *market sector* to derive market demand and to track customer properties, a *manufacturer sector* to model the very basic features a manufacturer's business model and to ensure a meaningful system's response to a market stimulus (this sector is actually instantiated for any assumed competitor type), and finally a *strategies & policy sector* to define the manufacturers' behaviour. The framework itself is implemented in Ventana Vensim 5.7 and is supported by various other tools in order to facilitate model configuration and scenario management.

4 An Exemplary Model Instantiation and Results

The framework's purpose is just to facilitate model formulation – it is neither a substitute for this process nor oblivious to case specific modifications. It provides a "minimum" simulation model, though, that incorporates the most important properties of a piracy competition. Therefore, it is but a starting point and must always be tailored to a certain situation. However, in this concluding section we do not want to investigate a specific case but rather want to show whether or not the model provides meaningful responses in piracy situations. The following short case for

instance will focus on why it is important to consider different market situations to evaluate a counterstrategy. Figure 2 shows the results of an exemplary simulation that compares the effects of a countermeasure ("secure distribution channels") dependent on two different market scenarios, which account for different levels of customers willing to buy counterfeits. The figure depicts the development of the installed base (sku=stock keeping unit) for each competitor type. Please note that the following description of the model is abbreviated - the actual parameters are idealised and set to "extreme" values for illustration purposes.

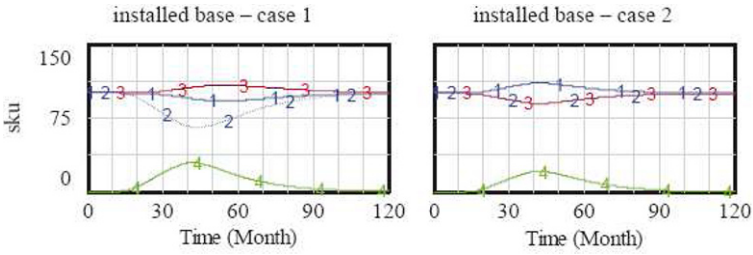


Fig. 2 Example model instantiation: simulation results for the installed base of the innovator, regular competitor and pirate (lines 2,3 and 4, respectively; line 1 refers to innovator but includes deceived customers)

The model was instantiated with three competitor types ("innovator", "regular competitor", "counterfeiter") sharing identical instantiations of their "manufacturer" sectors. Market demand is generated by customer reinvestments only (product life time=12 months). Since the counterfeiter is assumed to conduct only covered activities he has to "infiltrate" the innovator's distribution channels in order to generate demand – at which he succeeds with a certain probability. The counterfeiter enters the market in t=12 needing 12 months to achieve full operability, while the innovator starts securing his distribution channels at t=36 – also needing 12 months to expel the pirate completely. The buying decision of the customers is assumed to be based on a product's perceived performance as well as its price. The innovator's and competitor's products share the same properties, while the counterfeiter's products yield a much lesser performance. However, the model is configured such that the actual net utility of a counterfeiter's product equals the one of the innovator – given that a customer is able to identify the counterfeit as such. Thus, the counterfeiter offsets the lower performance with a much lower price. It is assumed that a customer, who is able to identify a counterfeit, is also willing to buy it – given he has access to an infiltrated distribution channel. The probability that a customer is buying counterfeits willingly is set to 0% and 100% in scenarios 1 and 2, respectively.

As the figure shows, the model behaves as expected: In case 1, the total market share for the innovator in terms of installed base (line 2) drops below the competitor's (line 3) – even if we account for deceived customers (line 1). Due to the lower performance of the counterfeits, every deceived customer will experience a lower net utility than expected. Thus, this will eventually impact on the overall market

expectations for the innovator's products. This results in a competitive advantage for the regular competitor – although the innovator's products actual performance still equals the one of the competitor. Consequently, if we reduce the amount of customers deceived, we actually reduce the impact of the counterfeits on market expectations for the innovator's products. In case 2, no customer is deceived any more, leading to equal market shares for both the competitor and innovator in terms of products sold. However, the total installed base relevant for the innovator (line 1), which includes the counterfeits, is significantly higher for the period in which counterfeits where actually sold. This is important to note and may be relevant for the innovator's strategy in situations where customers, who have bought a counterfeit product, may still demand for complementary goods provided by the innovator. In this case, it might have been wise to forgo any countermeasure and do nothing.

5 Summary and Outlook

Product piracy is still rampaging world's global markets and has spread into nearly every industry domain. Today, it is an essential business risk to be reckoned with and thus must find its way in strategic risk management. However, whenever the evaluation of a piracy situation and the formulation of fitting counterstrategies are concerned, research has yet to provide strategic decision making tools that go beyond guidelines and mere strategy suggestions. These tools must account for the dynamic complexity of the phenomenon. Empirical and case study research may provide clues and even hint at possible counterstrategies but they are not appropriate for ex-ante decision making. In this article we have argued for SD as an appropriate tool and provided a first simulation framework. However, the framework's purpose is just to facilitate model formulation – it is neither a substitute for this process nor oblivious to case specific modifications. Further, although we have shown its principal applicability, the model is still to be refined and has yet to be evaluated against a "real" case. This is currently done in a research project at the Fraunhofer ISI.

References

1. Derek Bosworth. *Counterfeiting and Piracy: The state of the art.* Intellectual Property Research Centre at St. Peters College, Oxford, 2006.
2. Julio O De Castro, David B. Balkin, and Dean A. Shepherd. Can entrepreneurial firms benefit from product piracy? *Journal of Business Venturing*, 23(1): 75–90, 2008.
3. Thorsten Staake, Frederic Thiesse, and Elgar Fleisch. The emergence of counterfeit trade: a literature review. *European Journal of Marketing*, 43(3–4): 320–349, 2009.
4. VDMA. *Produkt- und Markenpiraterie in der Investitionsgüterindustrie 2008.* Verband deutscher Maschinen- und Anlagenbauer (VDMA), Frankfurt a. M., 2008.

Microscopic Pedestrian Simulations: From Passenger Exchange Times to Regional Evacuation

Gerta Köster, Dirk Hartmann, and Wolfram Klein

Abstract Pedestrian dynamics play an important role in diverse fields of application such as optimizing traffic logistics, e.g. the optimization of passenger exchange times, or egress planning of buildings, infrastructures and even whole regions. Quantitative predictions of pedestrian dynamics, namely of egress times, is an essential part of optimizing pedestrian flows. To obtain quantitative results simulations must be as realistic as possible. Here, we present a new microscopic pedestrian simulator based on a cellular automaton. It reduces discretization artifacts as typically observed in cellular automaton models to a minimum without loosing their efficiency. It reliably captures typical crowd phenomena and can simulate up to 50000 pedestrians in real time.

1 The Problem: Efficient Management of Large Crowds

In the age of urbanization and mass events it has become crucial to efficiently manage crowds - both to ensure smooth transport and safety for everybody. A multitude of optimization problems emerge from these goals and are tackled by scientists from a very varied background with quite different tools.

The authors of this paper are mainly concerned with developing a reliable microscopic model of crowd movement that will provide insight in how certain scenarios, such as an evacuation of large crowds, evolve in time and space. The long-term goal is a practical tool to improve the overall situation in these scenarios where specific

Gerta Köster
University of Applied Sciences, Lothstr. 64, D-80335 München e-mail: gerta.koester@hm.edu

Dirk Hartmann
Siemens AG, Corporate Technology, D-80200 München e-mail: hartmann.dirk@siemens.com

Wolfram Klein
Siemens AG, Corporate Technology, D-80200 München e-mail: wolfram.klein@siemens.com

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_90,

goals vary: Reducing risk at a large event by removing congestion, reducing crowd density at hot spots, minimizing evacuation times or maximizing throughput at a train station are typical examples for optimization tasks. These applications lead to conflicting requirements: For one, model should be close to realistic and complex pedestrian behavior and at the same time the simulation should run sufficiently fast.

The simulation tool described in this paper allows to improve the value of possible optimization functions through the virtual experience that the user of the tool gathers. Alternatively, in an evacuation scenario, simulation experiments may be used to produce estimated upper bounds for evacuation time and to compare those with the lower bounds obtained with a network optimization algorithm [5, 7].

In Section 2 we outline the basics of the model that are responsible for its efficiency. In Section 3 we focus on strategies to reduce artifacts typical for cellular automatons [10, 12, 2, 1, 6] so that we can also quantitatively achieve a correct reproduction of movement nearly independent of the underlying discretizations. Section 4 deals with qualitative and quantitative tests that serve to build trust in the simulation results. The paper concludes with an outlook on future work.

2 The Simulation Model

The simulation model is a cellular automaton on a hexagonal grid (c.f. Fig. 1). That is, pedestrian dynamics are modeled on a microscopic level. At each time step each cell has a certain state: It is either empty or occupied by a pedestrian or a fixed obstacle. The positions are updated according to a set of rules.

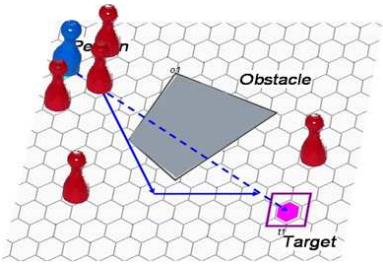


Fig. 1 Schematic sketch of a cellular automaton simulating microscopic pedestrian dynamics

Each pedestrian is generated with an individual free-flow velocity with which he or she moves in the absence of other persons. The assignment of the free-flow velocity is normally distributed following [11]. Pedestrians typically adjust their speed during the simulation, e.g. slow down with increasing crowd density. The update scheme of the cellular automaton has to guarantee, that pedestrians with a higher speed are allowed to move more often than pedestrians with a lower speed. In each time step of the simulation we decide which set of pedestrians is allowed to move: Faster persons are chosen more often so that, on average, pedestrians move with their prescribed velocity. Once the set of persons which are allowed to move is

determined, they are updated sequentially. That is, we must decide who moves first. We prefer a scheme that updates according to the "life time" of the persons in the scenario. This makes intuitively sense in a typical flow scenario, where people enter the observation area at some time and then leave it at another time.

The model's update rules are inspired by electro dynamics: pedestrians are attracted by targets modeled as long-range potentials $\phi_{\text{navigation}}(x,y)$, the navigation field [4]. They repel each other by a force modeled through short-range potentials $\phi_{\text{pedestrian}}(x,y)$. In the simplest case, the movement rule for a single pedestrian is purely deterministic: Find the unoccupied neighboring cell with the minimal total potential value $\phi_{\text{total}} = \phi_{\text{navigation}}(x,y) + \phi_{\text{pedestrian}}(x,y)$. The model is quite similar to the so-called static floor field cellular automaton (cf.[2] and further references in [9]). However, the artifacts caused by the discretizations of the cellular automatons lattice are reduced to a minimum.

2.1 The Attractive Navigation Potential and Repulsive Pedestrian Potential

The navigation potential $\phi_{\text{navigation}}(x,y)$ according to which persons move towards a target should be a scalar function that grows with increasing distance from the target. Several variants to construct appropriate navigation potentials can be found in the literature (for a summary see [4]). The most popular are navigation potentials based on 1-norms and navigation potentials using true Euclidean distances. Considering a 1-norm navigation potential $\phi_{\text{Dijkstra}}(x,y)$, the value of the navigation field scales with the number of discrete steps required to reach the target: given a set of cells, which are n steps away from the target, the distance of their neighbors is determined by simply adding one step. These methods are typically referred to as flood filling methods alluding to the analogy to the propagation of floods.

In our model we choose a much more precise variant of the flood filling methods: the Fast Marching Method (FFM). The idea is to consider the evolution of a wave front traveling with normal speed 1 rather than adding steps [4]. Mathematically the arrival times $T(x,y)$ of an evolving wave front is described by the Eikonal equation: $(\partial_x T(x,y))^2 + (\partial_y T(x,y))^2 = 1$, with $T(x,y) = 0$ for (x,y) on the initial curve $\mathbf{g}$, where the wave starts. The arrival times correspond to shortest distances with respect to the Euclidean metric since the wave evolves with a normal speed of 1, i.e. $\phi_{\text{navigation}}(x,y) = T(x,y)$. The direction of the shortest path is given by the gradient ∇T. By choosing a realistic measure for distances like the natural Euclidean metric artifacts typically found in navigation strategies based on 1-norms are avoided to a large extent.

Pedestrians repulse each other. This is also modeled with a potential $\phi_{\text{pedestrian}}$ which is added to the navigation potential. This additive treatment might yield local minima in the total potential ϕ_{total}. However, the potential $\phi_{\text{pedestrian}}$ changes continuously with the pedestrian movements. This makes dead locks unlikely. Since

a single persons's potential has a limited range, it is sufficient to only take into account pedestrians in a closer neighborhood (c.f. Fig. 2).

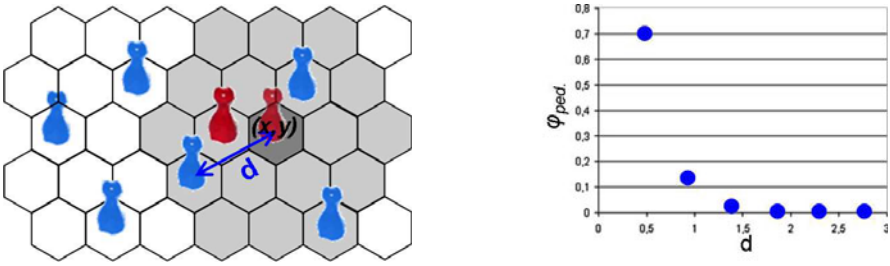


Fig. 2 Schematic presentation of the pedestrian potential calculation at the position (x,y) . Only the pedestrians on grey cells are taken into account. The corresponding potential field depending on the distance is plotted on the right side.

3 Strategies to Reduce Artifacts

The cellular automaton approach entails a somewhat unnatural spatial discretization and, through the update in time, a discretization in time. Both introduce artifacts in the pedestrian behavior such as a preference for routes in the direction of the symmetry axes of the discretization even with navigation fields based on true Euclidean distances. We introduce corrections within the movement strategies such that paths and travel times of virtual pedestrians no longer reflect the artifacts. The resulting movement can hardly be distinguished from continuum approaches that typically require significantly more computational effort. Here, we will focus on the reduction of time artifacts. For a detailed treatment of spatial artifacts we refer to [4].

Pedestrians in a cellular automaton are forced to move on the grid, Fig. 3. A pedestrian moving from cell A to cell B cannot move the direct way (red line), but has to take a detour (e.g. blue line). The length of the detour depends on the angle ω of the deviation from the direct path. The ratio of length of the detour to the direct path is $\alpha(\omega) = \cos\left(\frac{\pi}{6} - \omega\right) \cdot \frac{2}{\sqrt{3}}$. The detour is $\alpha(\omega)$ times longer than the direct path. Hence, the pedestrian must move faster to arrive at the same time at cell B as if he or she took the direct path. That is achieved by scaling the velocity by $\alpha(\omega)$.

4 Qualitative and Quantitative Test Scenarios

A meaningful comparison of very different approaches to the simulation of pedestrian crowds is best achieved by standard tests ensuring a minimum quality. In the absence of mandatory tests, we follow the suggestions of the RIMEA project [8], a

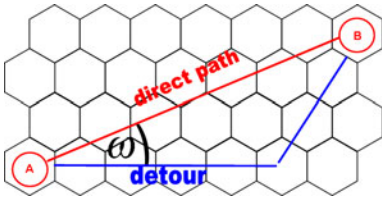


Fig. 3 Schematic explanation of the speed correction

society of enterprises and researchers in the field of pedestrian stream simulations. This initiative strives to ensure a minimum quality and comparability for simulation tools. The RIMEA project offers a number of qualitative tests as well as one quantitative test. Our simulation tools performs well in all cases ensuring the reliability we strive for. The quantitative tests demands that a simulation tool correctly capture the decrease of pedestrian speed with the increase of the crowd density – as given by Weidmann's fundamental diagram [11]. With our simulation tool, arbitrary fundamental diagrams can be obtained through calibration [3].

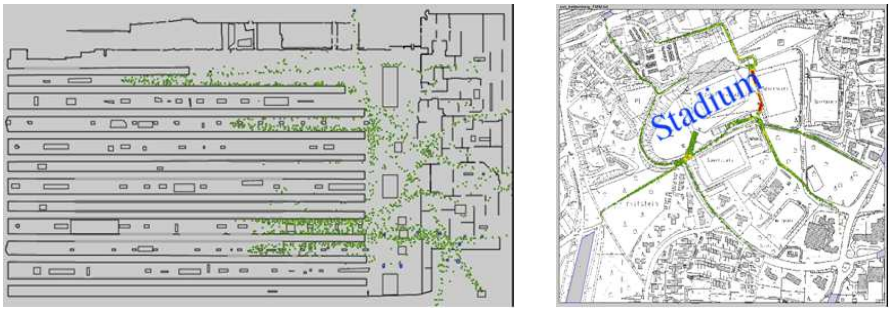


Fig. 4 Simulation of pedestrian traffic in a railway station and of a virtual regional evacuation of a stadium (Fritz Walter Stadium Kaiserslautern / Betzenberg). Pedestrian flows are reproduced in a natural way following the shortest path to the exits instead of following grid directions .

5 Discussion

We have presented an efficient cellular automaton model for pedestrian flow that at the same time reduces to a minimum artifacts typically inherent to cellular automaton approaches and still keeps their computational efficiency. It is able to correctly capture crowd phenomena and thus can be used as a basis for improvement of crowd management strategies, such as the optimization of egress times in an evacuation scenario.

In comparison to models that dispense with a spatial grid, such as social force models or some agent based models, the approach of a cellular automaton is some-

what simpler. This allows the use of highly efficient algorithms which are able to simulate up to 50,000 pedestrians in real time. It is generally assumed that the smaller the scale of the scenario, the bigger the influence of artifacts using cellular automata. By employing appropriate correction strategies discretization artifacts can be avoided to a large extent. However, some grid artifacts remain because each person occupies a full cell. It is unclear how small scale experiments, such as an isolated bottleneck, may be reproduced with sufficient quantitative precision. But so far, all our results suggest that the quality of the simulations is comparable to that of continuum approaches.

Furthermore, it would highly desirable to complement the RIMEA tests with scenarios based on measured data. However, collecting this data and finding appropriate measures for qualitative assessment of simulation results is still an open field of research. Most researchers, including the authors, are forced to rely on visual plausibility checks by comparing to, mostly unpublished, video data and intuition.

Acknowledgements The authors would like to thank the German Federal Ministry of Education and Research who funded our research through the priority program Schutz und Rettung von Menschen within the project REPKA - Regional Evacuation: Planning, Control and Adaptation.

References

1. V. Blue, M. Embrechts, and J. Adler. Cellular automata modeling of pedestrian movements. *IEEE Int. Conf. on Systems, Man and Cybernetic*, page 2320, 1997.
2. C. Burstedde, K. Klauck, A. Schadschneider, and J. Tittartz. Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. *Physika A*, 295: 507–525, 2001.
3. M. Davidich and G. Köster. Towards automatic and robust adjustment of human behavioral parameters in a pedestrian stream model to measured data. *PED2010*, 2010.
4. D. Hartmann. Adaptive pedestrian dynamics based on geodesics. *New Journal of Physics*, 12: 043032, 2010.
5. H. W. Hamacher, S. Heller, S. Ruzika, G. Köster, and W. Klein. A sandwich approach for evacuation time bounds. *PED 2010*, 2010.
6. H. Kluepfel. *A Cellular Automaton Model for Crowd Movement and Egress Simulation*. PhD thesis, 2003.
7. A. Kneidl, M. Thiemann, A. Borrmann, S. Ruzika, H. W. Hamacher, G. Köster, and E. Rank. Bidirectional coupling of macroscopic and microscopic approaches for pedestrian behavior prediction. *PED 2010*, 2010.
8. RIMEA. RIMEA e.V. Projekt. http://www.rimea.de, 2010.
9. A. Schadschneider, W. Klingsch, H. Klüpfel, T. Kretz, C. Rogsch, and A. Seyfried. Evacuation dynamics: Empirical results, modeling and applications. *Encyclopedia of Complexity and System Science, Robert A. Meyers (Ed.)*, 3: 3142, 2009.
10. A. Varas, M. D. Cornejo, D. Mainemer, B. Toledo, J. Rogan, V. Munoz, and J. A. Valdivia. Cellular automaton model for evacuation process with obstacles. *Physica A: Statistical Mechanics and its Applications*, 382(2): 631–642, 2007.
11. U. Weidmann. Transporttechnik für Fussgänger. em Schriftenreihe des IVT, 90, 1992.
12. K. Yamamoto, S. Kokubo, and K. Nishinari. Simulation for pedestrian dynamics by real-coded cellular automata. *Physica*, 379(2): 654, 2007.

Unexpected Positive Effects of Complexity on Performance in Multiple Criteria Setups

Stephan Leitner and Friederike Wall

Abstract The paper analyses the effect of organisational design elements on the level of achieved performance and speed of performance improvement in multiple criteria decision making setups. The analysis is based on an agent based simulation approach. Especially, we model multi criteria decision making as adaptive walk on multiple NK fitness landscapes with different levels of complexity. The results indicate that achieved performance and speed of performance improvement subtly depend on complexity and organisational design elements. The results might throw some new light on complexity. Complexity turns out to affect performance contribution and speed of performance improvement positively in some setups, with respect to multiple objectives.

1 Introduction, Research Question and Research Method

The phenomenon of complexity has been in interest of the organisational sciences since the early 1960's [11]. Many reasons for the fact that firms have to face increasing complexity have been discussed, but there are two especially prominent lines of explanation [10]. First, technological change and globalisation intensify competition and increase turbulence [5]. This contributes to the need of achieving multiple goals simultaneously. Second, pursuing multiple objectives and growing intraorganisational interdependencies cause a higher need of coordination at different levels of the firm [1].

Stephan Leitner
Alpen-Adria-Universitaet Klagenfurt, Dept. of Controlling and Strategic Management, Universitaetsstrasse 65-67, 9020 Klagenfurt, e-mail: stephan.leitner@uni-klu.ac.at

Friederike Wall
Alpen-Adria-Universitaet Klagenfurt, Dept. of Controlling and Strategic Management, Universitaetsstrasse 65-67, 9020 Klagenfurt e-mail: friederike.wall@uni-klu.ac.at

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_91,

Consequently, firms can respond with changes in organisational design in order to cope with the increase of complexity. Our research in particular analyses the effect of (1) organisational structure, (2) the coordination mode and (3) different incentive schemes on the achieved performance and the speed of performance improvements with respect to two objectives to be achieved simultaneously.

In order to investigate the research question, a research method that allows to map organisational design and different levels of complexity in multi criteria decision setups with different agents is needed. Following agency theory, agents' actions are oriented towards maximising the individual utility function. Consequently, these autonomous agents must be taken into account when choosing the research method. Due to the fact that repeated individual actions can lead to complex situations, the consideration of different agents would lead to intractable dimensions of formal modelling. Simulation, on the contrary, is a powerful method to face this complexity and analyse macro level complexities that result out of micro level interactions [9] and complexities that result out of different layers of decision making within a system [6]. The multitude of issues to be mapped would be particularly difficult to control in empirical research. Furthermore, variables may be contaminated because effects of variables under research cannot be disentangled by other effects [12]. Due to dimensions of formal modelling and limitations of empirical research, an agent based simulation approach appears appropriate [3].

2 Simulation Model

Our simulation model is based on the NK-model introduced by Kauffman [8, 7]. Especially, we model multi criteria decision making as adaptive walk on multiple NK fitness landscapes with different levels of complexity. In this context the notion of optimality has to be adapted from finding the (local) optimum at one landscape to finding good trade-offs among all objectives [2].

Our organisations have a binary ten-dimensional decision problem. Organisations have to make decisions $n^i \in N$ with $n^i \in \{0,1\}$ and $i = \{1,...,10\}$. All N decisions may intensely interact with each other or be completely independent from each other. These epistatic relations can be described by parameter K_z^i, which stands for the number of decisions n^j that affect the functioning of each performance contribution w_z^i additionally to decision n^i [4]. So, the performance contribution w_z^i to overall performance of that objective $z \in Z$ depends on the single decision n^i and eventually a number of other decisions n^j, with $0 \leq w_z^i \leq 1$, $i, j \in \{1,...,10\}$, $Z = \{1,2\}$ and $i \neq j$. For each objective $z \in Z$ overall performance W_z results as the normalised sum of performance contributions w_z^i, i.e.

$$W_z = \frac{1}{|N|} \sum_{i=1}^{|N|} w_z^i = \frac{1}{|N|} \sum_{i=1}^{|N|} f_z^i \left(n^i; n^{j=1}, ..., n^{j=K_z^i} \right) \tag{1}$$

where $i \neq j$.

The organisations consist of headquarters h and three departments $d \in D$ with various scopes of decision (two departments are in scope of three decisions and one department is in charge of four decisions). The set of departmental decisions are denoted as N^{own_d} while the set of decisions other departments are in scope of are denoted as N^{res_d}. In our model, we apply three different levels of complexity: (1) *low:* the decisions within a department interact with each other but are completely independent from the other departments' decisions (i.e. $K_z^i = 2$ for each w_z^i with $i \in N^{own_d}$ and $z \in Z$ for departments that are in scope of three decisions and $K_z^i = 3$ for each w_z^i with $i \in N^{own_d}$ and $z \in Z$ for the department that is in scope of four decisions), (2) *medium:* each department's decisions are partly independent from each other and they partly interact with other departments' decisions (i.e. $K_z^i = 4$ for each w_z^i with $i \in N^{own_d}$ and $z \in Z$) and (3) *high:* all decisions are fully interdependent (i.e. $K_z^i = 9$ for each w^i with $i \in N^{own_d}$ and $z \in Z$).

The departments operate under alternate incentive structures. Rewards depend on each objective's performance contributions of departmental decisions and the residual performance (i.e. the performance contribution of decisions other departments are in charge of) with varying weights denoted as $r_z^{own_d}$ and $r_z^{res_d}$. For simplicity we assume that with respect to multiple objectives Z for each department $d \in D$ utility U_d results as:

$$U_d = \sum_{z=1}^{|Z|} \left[r_z^{own_d} \left(\frac{1}{|N^{own_d}|} \sum_{i \in N^{own_d}} w_z^i \right) + r_z^{res_d} \left(\frac{1}{|N^{res_d}|} \sum_{i \in N^{res_d}} w_z^i \right) \right] \rightarrow \max! \quad (2)$$

While the departments' objective is to maximise rewards according to the incentive structure, the headquarters' objective is to maximise overall performance. For simplicity the headquarters' h utility U_h results as:

$$U_h = \left(\frac{1}{|Z|} \sum_{z=1}^{|Z|} W_z \right) \rightarrow \max! \quad (3)$$

Furthermore, we map two different coordination modes: (1) *central:* the departments make proposals (with respect to the decisions they are in scope of) to the headquarters which choose that proposal with the highest overall performance and (2) *decentral:* the departments decide autonomously about their partial decision problems. The overall configuration results as a concatenation of the departmental decisions without any intervention by the headquarters [10].

3 Results

Not surprisingly, we find that under the conditions of decentral as well as central coordination mode and equal weighted incentivisation increasing complexity leads to decreasing achieved overall performances (i.e. the headquarters' h utility U_h)

and decreasing speed these performances are achieved with. Pursuing two objectives with the same level of complexity leads to same levels of achieved performance and same speed of performance improvement per objective.

Conventional wisdom indicates that pursuing two objectives with different levels of complexity leads to a higher level of achieved performance for the less complex objective (lower K-values). Although performance contributions decrease with increasing complexity, our results show that in the *decentral decision making mode* (cf. table 1 panel A) the performance contribution of the more complex objective (higher K-values) is higher in cases *low/medium* or *medium/high* (with non overlapping confidence intervals). Pursuing objectives with level of complexity *low/high* leads to equal levels of performance for each two objectives. Results also indicate that, with respect to objective one and two, in case of the decentral decision making mode when pursuing objectives with different levels of complexity the performance improvement from period 0 to 1 (*speed 1*) is higher in case of the less complex objective but the performance improvement from period 0 to 10 (*speed 2*) is higher in case of the more complex objective (cf. table 1 panel A).

Table 1 achieved performances, incentivisation: $r_z^{own_d} = 1$ and $r_z^{res_d} = 1$ for each department $d \in D$ and each objective $z \in Z$.

complexity obj[a] 1 / obj 2	speed 1[b] obj 1	speed 1[b] obj 2	speed 2[c] obj 1	speed 2[c] obj 2	perf[d] obj 1	perf[d] obj 2	perf[e] overall
Panel A: decision making mode: decentral							
low / low	0.1295	0.1225	0.2166	0.2131	0.8966	0.8956	0.8961
low / medium	0.1127	0.0780	0.2016	0.2278	0.8796	0.8887	0.8842
low / high	0.0912	0.0415	0.1724	0.1915	0.8577	0.8581	0.8579
medium / medium	0.0714	0.0694	0.1973	0.1993	0.8553	0.8584	0.8569
medium / high	0.0592	0.0375	0.1667	0.1748	0.8288	0.8394	0.8341
high / high	0.0345	0.0344	0.1497	0.1467	0.8123	0.8110	0.8117
Panel B: decision making mode: central							
low / low	0.0846	0.0827	0.2180	0.2076	0.8978	0.8951	0.8965
low / medium	0.0744	0.1051	0.1879	0.2280	0.8740	0.8876	0.8808
low / high	0.0691	0.1280	0.1604	0.2135	0.8438	0.8634	0.8536
medium / medium	0.0954	0.0972	0.2011	0.2018	0.8553	0.8560	0.8557
medium / high	0.0907	0.1194	0.1757	0.1954	0.8240	0.8422	0.8331
high / high	0.1099	0.1098	0.1671	0.1670	0.8106	0.8107	0.8107

results are based on 9000 adaptive walks (each on 2 fitness landscapes).
[a] objective, [b] performance improvement from period 0 to 1, [c] performance improvement from period 0 to 10, [d] performance per objective normalised to respective maximum (after 100 periods), [e] overall performance normalised to respective maximum (after 100 periods), confidence intervals vary from 0.0019 to 0.0034 on the 99.9% level.

Pursuing objectives with different levels of complexity leads to similar results in the *central coordination mode* (cf. table 1 panel B). Achieved overall performance appears insensitive to coordination mode. In the central decision making mode all

combinations of objectives with different levels of complexity lead to a higher performance of that objective with the higher level of complexity (higher K-values).

Contrary to intuition, we find that in the central coordination mode when pursuing objectives with different levels of complexity *speed 1* and *speed 2* are higher in case of the more complex objective. In the decentral coordination mode for all levels of complexity *speed 1* is higher than in the central mode. While increasing complexity leads to decreasing *speed 1* in the decentral coordination mode, complexity has a positive effect on *speed 1* in the central mode. A higher level of complexity leads to decreasing *speed 2* in both modes.

Table 2 achieved performances, incentivisation: $r_z^{own_d} = 1$ and $r_z^{res_d} = 0.5$ for each department $d \in D$ and each objective $z \in Z$.

complexity obj[a] 1 / obj 2	speed 1[b] obj 1	speed 1[b] obj 2	speed 2[c] obj 1	speed 2[c] obj 2	perf[d] obj 1	perf[d] obj 2	perf[e] overall
Panel A: decision making mode: decentral							
low / low	0.1285	0.1251	0.2212	0.2080	0.8972	0.8910	0.8941
low / medium	0.1268	0.0728	0.2220	0.2056	0.9004	0.8666	0.8835
low / high	0.1147	0.0317	0.2069	0.1712	0.8941	0.8421	0.8681
medium / medium	0.0687	0.0675	0.1922	0.1924	0.8555	0.8587	0.8571
medium / high	0.0617	0.0333	0.1686	0.1547	0.8459	0.8329	0.8394
high / high	0.0321	0.0311	0.1279	0.1269	0.8164	0.8156	0.8160
Panel B: decision making mode: central							
low / low	0.0841	0.0841	0.2175	0.2154	0.9031	0.8966	0.8999
low / medium	0.0763	0.1047	0.1937	0.2200	0.8796	0.8776	0.8786
low / high	0.0699	0.1264	0.1646	0.2033	0.8463	0.8486	0.8475
medium / medium	0.0981	0.0953	0.1961	0.1953	0.8503	0.8506	0.8505
medium / high	0.0902	0.1169	0.1693	0.1867	0.8227	0.8338	0.8283
high / high	0.1110	0.1111	0.1674	0.1660	0.8066	0.8086	0.8076

results are based on 9000 adaptive walks (each on 2 fitness landscapes).
[a] objective, [b] performance improvement from period 0 to 1, [c] performance improvement from period 0 to 10, [d] performance per objective normalised to respective maximum (after 100 periods), [e] overall performance normalised to respective maximum (after 100 periods), confidence intervals vary from 0.0020 to 0.0035 on the 99.9% level.

Changing incentivisation to $r_z^{own_d} = 1$ and $r_z^{res_d} = 0.5$ for each department d and each objective z in the *decentral coordination mode* (cf. table 2 panel A) apparently does not substantially affect overall performances. But pursuing objectives with different levels of complexity then leads to higher relative performance contributions and higher *speed 1* and *speed 2* in case of the less complex objective.

In the *central coordination mode* (cf. table 2 panel B) changed incentivisation leads to equal levels of relative performance contributions in cases *low/medium* and *low/high*; levels of complexity *medium/high* lead to a higher relative performance contribution of that objective with the higher level of complexity. *Speed 1* and *speed 2* are higher in case of the more complex objective.

4 Implications and Conclusion

Conventional wisdom indicates that complexity affects the ease of achieving objectives negatively. We find that complexity in multiple criteria setups in certain situations affects relative performance contributions and speed of performance improvements positively.

The results indicate that relative performance contribution and speed of performance improvement can be affected by incentivisation. Especially, we show that putting weight on departmental performance in the incentive scheme leads to higher relative performance contributions of that goal with the more intense intra-departmental interactions.

The results show that the level of achieved performances and the speed of performance improvement subtly depend on complexity and organisational design elements. Thus, apparently corporate planning in multi criteria setups intensely should consider organisational design. Researchers are called to replicate presented results and study the observed phenomenon in depth.

References

1. Chrisopher A. Bartlett and Sumantra Goshal *Managing Across Borders: The Transnational Solution*. Harvard Business School Press, Boston, 1989.
2. Carlos A. Coello Coello, Artura H. Aguirre, and Eckart Zitzler. Evolutionary multi-objective optimization. *European Journal of Operational Research*, 181: 1617–1619, 2007.
3. Jason P. Davis, Kathleen M. Eisenhardt, and Christopher B. Bingham. Developing theory through simulation methods. *Academy of Management Review*, 32(2): 480–499, 2007.
4. Koen Frenken. A fitness landscape approach to technological complexity, modularity and vertical disintegration. *Structural Change and Economic Dynamics*, 17: 288–305, 2006.
5. Gary Hammel and Coimbatore K. Prahalad. Competing for the Future. *Harvard Business Review*, 72(4): 122–128, July-August 1994.
6. Mohsen Jahangirian, Tillal Eldabi, Aisha Naseer, Lampros K. Stergioulas, and Terry Young. Simulation in manufacturing and business: A review. *European Journal of Operational Research*, 203(1): 1–13, 2010.
7. S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theorecital Biology*, 128(1): 11–45, September 1987.
8. Stuart A. Kauffman. *The Origins of Order. Self-organization and Selection in Evolution*. Oxford University Press, Oxford, 1993.
9. Tieju Ma and Yoshiteru Nakamori. Agent-based modeling on technological innovation as an evolutionary process. *European Journal of Operational Research*, 166(3): 741–755, 2005.
10. Nicolaj Siggelkow and Jan W. Rivkin. Speed and Search: Designing Organizations for Turbulence and Complexity. *Organization Science*, 16(2): 101–122, March-April 2005.
11. Herbert A. Simon. The Architecture of Complexity. In *Proceedings of the American Philosophical Society*, volume 106, pages 467–482, December 1962.
12. Geoffrey B. Sprinkle. Perspectives on experimental research in managerial accounting. *Accounting, Organizations and Society*, 28(2–3): 287–318, 2003.

Towards a Methodical Synthesis of Innovation System Modeling

Silvia Ulli-Beer and Alexander Wokaun

Abstract The following proposition for a methodical synthesis of innovation system modeling specifically refers to the scientific task of theory building and refinement based on case study research. It draws on innovation system research, system dynamics modeling and combines Beer's deliberations on scientific modeling. More specifically we derive proposition, concerning what modeling tactics of system dynamicists are most fruitful to contribute to theory building and refinement within the field of innovations system and policy research.

1 Introduction

In the last ten years system approaches to innovation and policy research have been increasingly adopted by different agencies at different levels (e.g. the OECD, the European Commission, or the European Environment Agency). Scholars of these approaches (e.g. [4, 11, 9, 7]) aim at offering a complementary analysis framework to economic mainly market failure based studies on innovation and technology change. The interest in the feedback mechanisms and the functional dynamics of innovation system approaches creates an opportunity for system dynamics researchers. They are applying a scholarly developed modeling approach that aims at identifying the structure and processes underlying the behavior patterns of dynamical complex systems. The application of such computer based analysis frameworks would enhance the progress of innovation system approaches in theory building and in innovation policy making. Modeling exercises would help to explore how sensitive coupled innovation systems are concerning policy action or inaction as well as to identify most

Silvia Ulli-Beer
Paul Scherrer Institut, Villigen, Switzerland, e-mail: silvia.ulli-beer@psi.ch

Alexander Wokaun
Paul Scherrer Institut, Villigen, Switzerland, e-mail: alexander.wokaun@psi.ch

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_92,

effective policy packages. A well grounded framework as discussed in this paper is seen as important and helpful since it will facilitate the confidence building process between researchers from different communities and disciplines.

2 Towards a Research Strategy Framework

We introduce a research strategy that has been inspired by Beer's methodology of topological maps and scientific modeling [3] and has been applied in different studies on dynamics of innovative systems (e.g. [15]). Stafford Beer distinguishes in his framework on scientific modeling the managerial situation and the scientific situation (compare figure 1). By definition the perception and data reduction process in the managerial situation is systematically unrecognized whereas in the scientific situation the reduction process is deliberate and creates a well understood detail of a scientific view. He describes the process of systemic modeling as mapping of the managerial concept model into a scientific concept model in such a way that the structure is preserved.

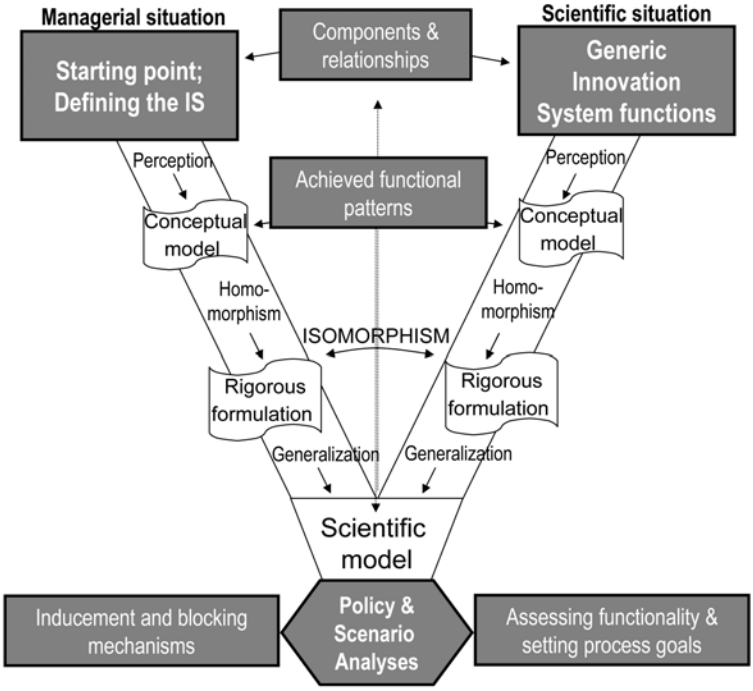


Fig. 1 Illustration of a reconciled research frame combining Beer's methodology of topological maps and scientific modeling with the scheme for technological innovation system analysis [3, 4].

3 Theses on Requirements for Innovation System Modeling

Based on this framework four propositions are derived, concerning what modeling tactics of system dynamicists are critical for a fruitful contribution to theory building and refinement within the field of innovations system and policy research. In innovation research the empirical phenomena in focus is often related to preliminary innovation concepts. But, different cases have their anomalies therefore each case needs to be carefully described if it should enhance theory building. Eisenhardt [6] emphasizes that in this phase the data gathering process may be very flexible, and that it may combine data collecting and analysis at the same time. At this stage, also "personal theories" or practitioner's accounts of the problem situation are an important data source. In a messy problem situation collaborative research methods may be most useful in order to define what exactly the relevant problem is for the involved practitioners (c.p. [13]). Without a clear description of the phenomena under study, the subsequent theory building process may be difficult to follow by other researchers working on innovation policy issues (c.p. [5]). It provides the common starting ground for a dialogue with the thematically interested scientific community. System dynamics has developed helpful methods or scripts that gear the focus towards dynamical characteristics of the phenomenon. The use of longitudinal graphs as reference modes of a dynamical problem is a typical example (e.g., [14, 1]). The first step "defining the problem situation" builds the basis for choosing the most adequate scientific framework. It actually answers the question "What is the object of research or the unit of analysis?" and does not yet contribute to answering questions of theorizing such as how, when and why (c.p. [2]).

1. Proposition Specifying the problem situation: *The problem definition is mainly an empirical task and serves as basis for the dialogue with thematically interested innovation policy researchers. It requires a detailed account on the empirical phenomena under focus that tries to push thinking beyond theoretical lenses.*
The outcome would be an empirical case description

Acknowledging the assumption made by Ferraro, Pfeffer et al. [8] that the acceptance of scientific theories does also depend on the political and rhetorical skill, and power of the proponents and opponents it becomes important to identify the relevant cognitive community in the field of innovation research. Since the innovation policy research field has developed its own evaluation communities based on a merit-based market system, it becomes crucial to identify early on the appropriate thematic audience and their communication system (i.e. journals and conferences). Different research objects and levels of analysis may point to different target audiences and publication outlets. Identifying the thematic reference-authorities also helps to identify corresponding and helpful research perspectives and approaches and frameworks. Finally it may become crucial for identifying the dialogue partners and for legitimating the own ongoing research.

2. Proposition Selecting the thematic reference-authorities: *Increasing expertise in innovation system modeling is an applied modeling discipline that should be grounded in a thematically field. It requires that its contributions to theory development and refinement can be discussed and judged within innovation system communities and researchers of the broader innovation and policy research field.*
The outcome would be a tentative communication and publication plan

The empirical case description and first speculations on a conceptual model from a practitioners view guides the identification of helpful scientific perspectives and frameworks from previous research on similar cases. The comparison of emergent concepts, hypotheses or theory with the extant literature is an essential feature of theory building. An important aspect in this step is to discuss case selection based on the concept of theoretical sampling. How adequate is the case in order to replicate or extend an emergent theory. This first evaluation helps to define the limits early on for generalizing the findings and to identify its relevance. The theoretical positioning results in the identification of the research gap based on an evaluation of previous explanation frames. The evaluation process helps to answer the question how adequately previous theorizing may explain the observed phenomena and where the limitations are. Subsequently, it can be speculated where the new case may contribute to an increased understanding and to informed policy making. Also, this evaluation is guiding the formulation of alternative (dynamic) hypotheses and the actual research questions.

3. Proposition Discussing tentative reference frameworks: *System dynamics modeling can enhance innovation system research by providing scientific tools for virtual experimentation and computer assisted theory building that would be otherwise impossible. It requires the identification of corresponding theorizing starting points within the broader innovation policy research field in an early conceptualization phase. It requires further the theoretical discussion of case sampling and the dynamic hypothesis regarding its likelihood to replicate or extend emergent theory within the conceptualization phase.*
The outcome would be a clear definition of the scientific modeling task with formulated research questions and scientific sound dynamic hypotheses including the definitions on time scale, boundary and levels of analysis.

Ongoing research on system dynamics modeling is highlighting the importance of empirical rigor and is pointing to traditional theory building research such as grounded theory approaches or case study research that are most helpful for supporting model conceptualization [12]. They also point out the tension between empirical rigor and parsimony of the resulting theory. In addition, the challenge of empiricism and system dynamics modeling needs to be addressed. The efforts and cost of computer assisted theory building can become overwhelming if the empirical data collection and analysis methods are applied as comprehensively as in traditional theory building approaches and if modeling manly becomes an additional specification. In the traditional theory building literature different tactics and methods are

highlighted that provide stronger substantive of hypotheses and their internal validity (e.g. case study write ups, triangulation by multiple data collection and analysis methods). However, the challenge is not to become highly competent as social scientist in the first phase and then as system dynamics modeler in the second, but to develop expertise in computer assisted theory building on innovation research that is optimally effective and efficient. Therefore tailored empiricism for computer assisted theory building needs to be developed that meets the critical requirement of scientific modeling and scientific theory building. Such ambitions become an important research avenue for applied modeling projects in general. Research into synergies and redundancies between traditional social science approaches and computer assisted approaches for theory building needs to be identified. Also, further research in so called "small models" [10] that address the challenge of homomorphism and parsimony of scientific modeling may become most helpful in complex innovation policy tasks.

4. Proposition Developing expertise in innovation system modeling: *Innovation system modeling expertise depends on tailored methods of computer assisted theory building. It requires a methodical synthesis of innovation system empiricism and modeling that is guiding the development of small models for innovation policy.* ***The outcome would be a parsimonious model with high explanatory power on innovation dynamics that allows conducting virtual policy experiments.***

4 Discussion and Conclusions

The propositions are not contradicting any theorizing in the field of system dynamics on how the concepts of feedback loops, rates and stock variables should be used to construct models on social phenomena. But they reflect on critical requirements for positioning this scholarship as an applied modeling discipline in the field of innovation policy research. The main argument is that for the scientific legitimation of endogenous simulation models on induced technology change, the thematically oriented community of the innovation research field becomes as important as the methodological focused system dynamics community. In addition, it highlights that expertise in innovation system modeling requires a tailored methodological synthesis that supports both the empirical identification and the scientific mapping of the discriminating structures of the specific behavioral phenomena of an innovation system. Empirical research methods should facilitate the exploration of the link between practice, theory and modeling. At the best such a methodical synthesis may be guided by general innovation frameworks (e.g. the multilevel perspective on transition) or core concepts of innovation system analysis as suggested by Bergek, Jacobsson et al. [4]. This implies that research also should reflect on (less) successful method mixing experiences in order to improve expertise on innovation system modeling.

Acknowledgements The authors are grateful for helpful comments from Professor Ruth Kaufmann-Hayoz and technical support from PhD-candidate Matthias O. Müller on the paper. It grew out of several projects that were financially supported by the Swiss National Science Foundation, novatlantis from the ETH-domain, the Swiss Federal Office of Energy, and the Competence Center of Energy and Mobility at Paul Scherrer Institut as well as the municipal building departement of Zürich. Many discussions with research collaborators, in particular with Matthias O. Müller, Stefan N. Grösser, Stephan Walter and Manuel Bouza as well as Susanne Bruppacher, stimulated the presented research.

References

1. D. F. Andersen and G. P. Richardson. Scripts for group model building. *System Dynamics Review*, 13(2): 107–129, 1997.
2. S. B. Bacharach. Organizational theories: Some criteria for evaluation. *The academic management review*, 14(4): 496–515, 1989.
3. S. Beer. The viable system model: Its provenance, development, methodology and pathology. *Journal of the Operational Research Society*, 35(1): 9, 1984.
4. A. Bergek, S. Jacobsson, B. Carlsson, S. Lindmark, and A. Rickne. Analyzing the functional dynamics of technological innovation systems: A scheme of analysis. *Research Policy*, 37: 407–429, 2008.
5. P. R. Carlile and C. M. Christensen. The cycles of theory building in management research (version 6.0). *Available from http://www. innosight. com/documents*, Accessed May 19 2010, 2005.
6. K. Eisenhardt. Building theories from case study research. *Academy of Management Review*, 14: 532–550, 1989.
7. J. Fagerberg. Innovation studies – the emerging structure of a new scientific field. *Research Policy*, 38: 218–233, 2009.
8. F. Ferraro, J. Pfeffer, and R. I. Sutton. How and why theories matter: a comment on Felin and Foss (2009). *Organization Science*, 20(3): 669–675, 2009.
9. F. W. Geels, M. P. Hekkert, and S. Jacobsson. The dynamics of sustainable innovation journeys. *Technological Analysis & Strategic Management*, 20(5): 521–536, 2008.
10. N. Ghaffarzadegan, J. Lyneis, and G. P. Richardson. Why and how small system dynamics models can help policymakers: A review of two public policy models. *Proceedings of the 26th International Conference of the System Dynamics Society, July 26-31, Albuquerque, NM USA*, 2009.
11. S. Jacobsson. The emergence and troubled growth of a 'biopower' innovation system in Sweden. *Energy Policy*, 36(4): 1491–1508, 2008.
12. B. Kopainsky and L. F. Luna-Reyes. Closing the loop: Promoting synergies with other theory building approaches to improve system dynamics practice. *Systems Research and Behavioral Science*, 25: 471–486, 2008.
13. M. Mueller, S. Ulli-Beer, and S. Groesser. How do we know whom to include in collaborative research? Towards a method for the identification of experts. *European journal of operational research*, submitted, 2010.
14. J.D. Sterman. *Business Dynamics. Systems Thinking and Modeling for a Complex World.* Irwin McGraw-Hill, Boston, 2000.
15. S. Ulli-Beer, S. Bruppacher, S. Grösser, S. Geisshüsler, M. O. Müller, M. Mojtahedzadeh, M. Schwaninger, F. Ackermann, D. Andersen, G. Richardson, R. Stulz, and R. Kaufmann-Hayoz. Introducing an action science venture: Understanding and accelerating the diffusion process of energy-efficient buildings. In *Proceedings of the 24th International Conference of the System Dynamics Society, (23–27 July 2006)*, Nijmegen NL, 2006.

III.6 OR in Life Sciences and Education – Trends, History and Ethics

Chair: Prof. Dr. Gerhard Wilhelm Weber (METU Ankara, Turkey)

That this new Selected Topic of "OR in Life Sciences and Education – Trends, History and Ethics" is given, reflects both the rapid changes which the modern world is experiencing, as challenges and chances, and that our OR has responded to these developments by offering scientific hospitality to researchers and practioneers, by providing and refining the methods needed. The foundation of corresponding EURO Working Groups and of working groups of national OR societies has been milestones on this way. Those changes are sometimes associated with labels such as, e.g., global warming, financial crisis, globalization, swine flu, and there are the UN Millennium Development Goals, especially, to fight poverty.

Our new Selected Topic invites colleagues from all disciplinary backgrounds to meet, to present and further discuss how modern, interdisciplinary OR with its quantitative methods contributes to urgent problems of our societies, of nature ("bio") and environment, with a special emphasis on the "human factor", the "social factor", on the improvement of living conditions and on ethics.

Operations Research at the Swiss Military Academy: Supporting Military Decisions in the 21st Century

Peter T. Baltes, Mauro Mantovani, and Beat Suter

Abstract At the beginning of the 21st century, there is a general consensus that most military operations have experienced a fundamental shift towards supporting the efforts of civil authorities. Consequently, research and curricula of institutions devoted to the training of military personnel must reflect these changes. This research illustrates the respective implications for the case of the Swiss Military Academy. In particular, it highlights the joint effort of two chairs (Economics of Defence/Strategic Studies) in the field of Operations Research/Military Decision Theory.

- Mauro MANTOVANI develops a model comprising the preconditions necessary for the reemergence of a state-to-state conflict with military means within Europe, which, in principle, allows for early-warning against upcoming inter-state military conflicts in general.
- Peter T. BALTES and Beat SUTER investigate if the Real Option Approach developed in Investment and Finance can provide an answer on when to rely on a surge as well as how to design it.

Peter T. Baltes
Swiss Military Academy at ETH Zurich, Kaserne, 8903 Birmensdorf e-mail: peter.baltes.bp@vtg.admin.ch

Mauro Mantovani
Swiss Military Academy at ETH Zurich, Kaserne, 8903 Birmensdorf e-mail: mauro.mantovani@vtg.admin.ch

Beat Suter
Swiss Military Academy at ETH Zurich, Kaserne, 8903 Birmensdorf e-mail: beat.suter.sb@vtg.admin.ch

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 591
DOI 10.1007/978-3-642-20009-0_93,

1 An Early-Warning Model for the Reemergence of a State (to State) Military Aggression

Acknowledgements Mauro Mantovani

The Government's 2010 report to Parliament on the security of Switzerland considers the risk of a military conflict inside an area directly affecting Switzerland to be very small within the foreseeable future, but it does not categorically exclude this contingency for all times. Accordingly, the Swiss armed forces continue to be tasked to allocate resources to maintaining the capabilities required in conventional defence scenarios. Classical and autonomous defence is called the core competence of the Swiss armed forces. However, there has not been given much thought in Switzerland neither to measuring the likelihood of such a development nor on the way to militarily react to it – except for a rather hazy concept of military surge ("Aufwuchs" - see second contribution).

The model presented here tries to identify the preconditions for the reemergence of a state (to state) military conflict within the area affecting Switzerland directly, i.e. Europe, but it is applicable to inter-state military conflicts in general. It is supposed to provide the career officers and future military leaders with a set of interrelated and measurable factors enabling them to give an early warning notice to the policy makers, based on thresholds they are supposed to define themselves. The model is

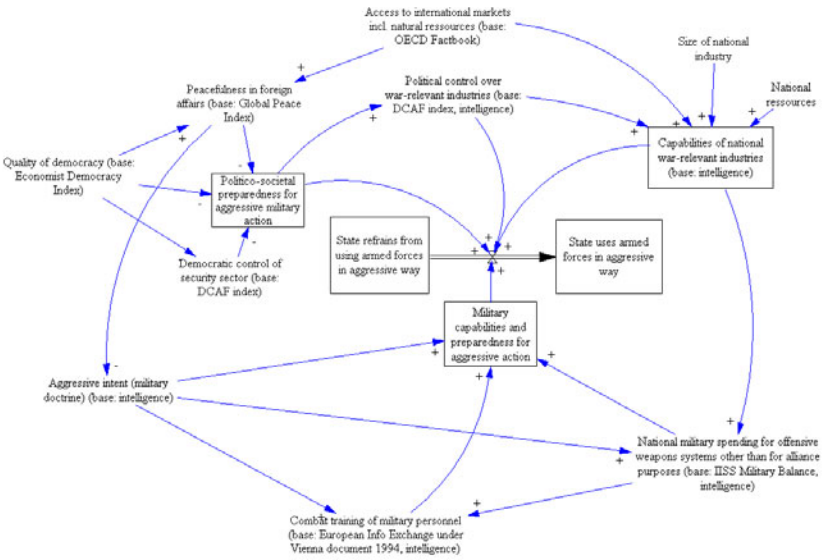


Fig. 1 An Early-Warning Model for the Reemergence of a State (to State) Military Aggression

primarily based on Andrea Riemer's [8] "multi-methodical early-warning approach" comprising "core variables", interlinked with signals of graded intensity, constituting clusters, so called "core-variable complexes". Further, the model's conceptual framework takes different approaches into account. First, it refers to the advanced game theory of [2] and [1] in that the aggressive state is understood as interacting with its victim's behaviour. Second, since the model argues on two levels (national level variables and intrastate variables) it rests on some assumptions of [5, 6]'s neoliberalism. And third, for the evaluation of the core variables we used the approach chosen by [10].

Our model contains three such core variables: the "Military capabilities and preparedness for aggressive action", the "Politico-societal preparedness for aggressive military action" and, thirdly, the "Capabilities of national war-relevant industries". As to the military core variable, it depends highly on an aggressive intent as well as on the level of combat training of the military personnel and on the level of military spending. Each of these variables needs special refinement according to historical experience, e.g. an aggressive intent shall always be directed against a *specific* neighbour or the military capabilities shall need to reach a *certain* degree of superiority over those of that neighbour. Obviously it also matters whether a state increases its military spending for alliance purposes or for its own national defence purposes.

The second core variable, the political and societal inclination of a state to become aggressive, depends mainly on the quality of its democracy and of its peacefulness (in the conduct of foreign affairs) – the underlying empirical hypothesis being, of course, that democracies are disinclined to wage war (especially against other democracies). In the description of our democratic variable we follow [4] and therefore accept that if democracy is to be measured participation must be considered. So we included participation as the democratic control over the security sector.

The third dimension is economic. Here, in our view, the degree of integration of an economy is the critical point of departure: The more access an economy has to international markets, the higher the nation's prosperity – with an ambivalent effect: on the one hand, it will increase the peacefulness of that nation because that state does not need to literally conquer these markets and resources and because its foreign assets are themselves vulnerable to aggression. On the other hand, prosperity offers the preconditions for increasing the capabilities of that state's "war-relevant industries". This third core variable is directly influenced by the level of political control over war relevant industries – on the assumption that a war-willing state cannot allow civil or even foreign interference into its prime base for the build-up and sustained functioning of its military force.

All these variables are measured with independent data, drawn in part from (freely or commercially) available indices. But since we are dealing with a highly sensitive national security issue, all these public indices need to be confirmed by intelligence data. For some variables like the capabilities of the war-relevant industries, we even depend completely on the assessment of our intelligence services. And of course, the variables' effect on each other needs to be carefully weighed.

The purpose of this model, as it stands now, is mainly didactical: It aims firstly at teaching the techniques of modeling and addressing recurring problems of modeling such as measuring factors or defining critical thresholds. Second it allows a structured analysis of a key security policy problem in general. By using this model we are able to identify the relevant factors, weigh their impact and clarify their interdependence. Thirdly, by means of this model, the military are supposed to reflect their role while a conventional military threat evolves. Key questions are their possibilities to influence this process from the outside or the appropriate options the military could submit to the political decision makers – for example: general mobilization, "symmetric" or "asymmetric" surge, appeasement, joining an alliance etc.
For Switzerland it means that if the country were serious about the assumption that a threat of conventional war could reemerge in Europe one day, it would be consequent to further develop this model by adjusting the relevant factors, by feeding the model with real world data on a permanent basis and by conducting simulations. The Military Academy is busy advancing the refinement of this model, drawing from historical cases of escalating conflicts. Eventually, this will allow the model to be used as a truly operational early-warning tool.

2 Real Options: A Tool for Surge Decisions?

Acknowledgements Peter T. Baltes / Beat Suter

At the beginning of the 21st century, a conventional conflict in Europe seems highly unlikely. But the horrific experiences of former Yugoslavia during the 1990s remind us that a return of history can always happen; and it can happen fast. The combination of a current low probability of military crises' – offering a tempting reason to ask for further reductions in defence budgets – and the uncertainty about future developments (at least in the mid-term perspective) leads to the following question: What levels of capacities are *now* required for the Armed Forces – in order to make sure that they can counter the possible threats of *tomorrow*?
In reaction to this challenge, the Swiss Army developed a so-called surge concept in 2006 – see [3]. Being conceptualized as a first stepping stone, the corresponding draft remained silent in regard to the provision of operational recommendations: What capacities should be classified as core skills that need to be preserved and developed over time because of their importance regarding future threats? How can it be ensured that credible threat signals will lead to a consistent coordination effort by all key players – the politicians, the military, the intelligence service, the economy etc.? What do the timing structure and the cost structure of a convincing surge reaction look like?
When it comes to building up a theoretical foundation for the surge decision from an economic perspective, the so-called Real Options Approach (ROA) seems particularly suited for this task: By adapting models of option pricing employed in Finance, the ROA (developed in the late 1980s / early 1990s) strives to overcome a major problem of the classical method of "Net Present Value": Designing flexible

investment strategies for a dynamic environment – see, for example, [7] and [9].
The paper gives a first glance of what might be expected from the transfer of the ROA method to military decisions by investigating the fundamental question(s) of capacity planning: When should a doctrine be preferred that seeks (mostly for cost-cutting reasons) to synchronize threats with capacities? When is a rather inflexible approach – deploying a maximum number of military units – superior because it is able to deliver a performance that comes close to "playing safe" in national security?
A simple model explores the ficticious case of a conference protection versus different types of threats. In particular, the analysis focuses on the trade-offs between some of the determinants – such as the probability that no escalation will take place ($= p$), the costs of mobilization ($= k$) and the losses to society due to escalations in the face of insufficient force deployment ($= d$). By adopting the economic concept of the indifference curve, *Figure 2* illustrates the corresponding trade-offs: For example, the first solid line from left represents for $p = 90\%$ all the combinations of d and k where the decision maker should be *indifferent* between the surge option and the option to "play safe" by deploying the maximum number of units available. In general, high values of p, high values of k and low values of d increase the attractiveness of the surge option.
Consequently, this simple model is already able to explain why the Swiss Army has taken a leading role in Europe concerning the willingness to pursue the surge conception as a main cornerstone of the national military doctrine: In comparison to a professional army, the surge option is *ceteris paribus* better suited to a militia

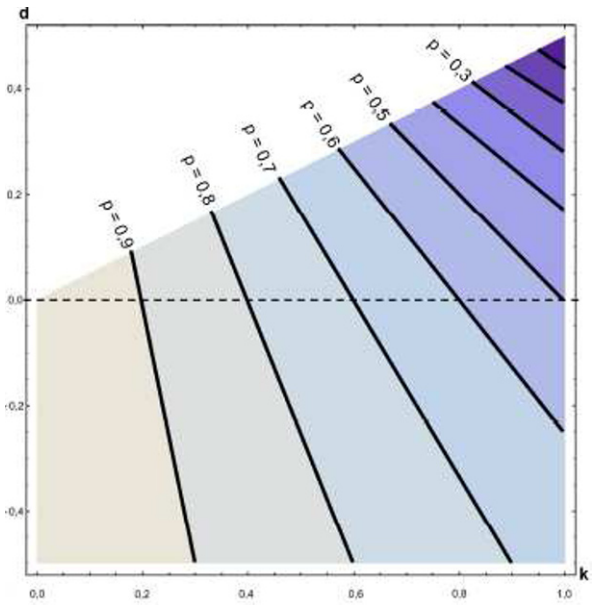


Fig. 2 Force Deployment in a Conference Setting - Surge versus "Playing Safe"

system because the latter faces much higher (social) costs of mobilization. In addition, because of its geostrategic position (being surrounded by friendly buffer states), Switzerland enjoys longer reaction times to most emerging threats than the bulk of other countries in Europe. In turn, this lowers the probability of being caught off-guard: $1 - p$.

In conclusion, the ROA provides a systematic perspective on how to incorporate flexibility in military decisions: For example, the different categories of real options identified by the literature should be incorporated into the general checklist of military decisions. However, this paper represents only a beginning. Further research has to be devoted to sorting out the similarities and the differences between the pricing of financial assets, the pricing of investment projects and the pricing of the national security.

References

1. R. Axelrod. *On Six Advances in Cooperation Theory*. Lucius & Lucius, Stuttgart, 2000.
2. R. Axelrod and R. Keohane. Achieving Cooperation under Anarchy: Strategies and Institutions. *World Politics*, 38(1): 226–254, 1985.
3. Bundesrat. *Botschaft über Änderungen der Armeeorganisation und des Bundesgesetzes über Massnahmen zur Verbesserung des Bundeshaushaltes (Rechtliche Anpassungen zur Umsetzung des Entwicklungsschrittes 2008/11 der Armee)*. Bundeskanzlei, Bern, 2006.
4. U. Marti. *Demokratie. Das uneingelöste Versprechen*. Rotpunktverlag, Zürich, 2006.
5. A. Moravcsik. Taking Preferences Seriously: A Liberal Theory of International Politics. *International Organization*, 51(4): 513–553, 1997.
6. A. Moravcsik. The European Constitutional Settlement. *World Economy*, 31(1): 158–183, 2008.
7. J. Mun. *Real Options Analysis Ű– Tools and Techniques for Valuing Strategic Investments and Decisions*. John Wiley & Sons, Hoboken, 2005.
8. A. Riemer. *Theorien internationaler Beziehungen und neue methodische Ansätze*. P. Lang, Frankfurt am Main, 2006.
9. R. Shockley. *An Applied Course in Real Options Valuation*. Thomson South-Western, Mason, 2005.
10. F. Vester. *Die Kunst vernetzt zu denken*. dtv Verlag, München, 2007.

Formalization of Models for the Analysis of the Phenomenon of Cross-Culture in a Multi-Ethnic Scholastic Environment

Rina Manuela Contini and Antonio Maturo

Abstract The phenomenon of interculture in the plural school [9] in the contemporary societies is dealt with. Interculture is very complex [2, 17, 20] and it can be only partially handled with the use of particular techniques of the operational research. The techniques used in this research are the Analytic Hierarchy Process (AHP), for the division of a complex objective in a hierarchical structure, assessment of scores, and ranking of alternative strategies [18, 19], statistic analysis made up of questionnaires administered to a high number of preadolescents attending the second and third year of first degree secondary school in Abruzzo (Italy), and research of the latent structures of phenomenon.

1 Interculture and Multi-Ethnic Scholastic Environment

International migrations and globalization processes deeply change Western societies [2, 3, 20]. The development of a multi-ethnic and multi-cultural character in such societies set the question of building an inter-ethnic living together [17, 6]. Particularly, "new generations" of immigrant origin, born or grown up in welcoming countries, require equal treatment and social promotion opportunities [10]. School is a fundamental institution for the acquisition of necessary competencies for economic and political integration of the citizens of tomorrow and to promote living together in multiple societies.

The importance of the school before migrations is highlighted considering the increase of foreign students in European scholastic systems. The school institution is

Rina Manuela Contini
Department of Social Sciences, University of Chieti-Pescara, via dei Vestini 31, 66100 Chieti, Italy.
e-mail: rm.contini@unich.it

Antonio Maturo
Department of Social Sciences, University of Chieti-Pescara, via dei Vestini 31, 66100 Chieti, Italy.
e-mail: amaturo@unich.it

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_94,

characterized by a plural experience: in the classes there are multiple languages, cultural backgrounds and different levels. In such a context it is important to consider the point of view of intercultural education, that is seen as central within European trends and the Italian regulation [9].

Intercultural teaching is aimed at promoting the ability of knowing, appreciating and respecting cultural diversity and, at the same time, to "the research of social cohesion, in a new vision of citizenship suitable for current pluralism in which most of the attention is focused on building a convergence towards common values" ([11]: 142-143).

2 The Research of Critical Variables in a Intercultural School from the Teacher's Point of View

The complex phenomenon of intercultural integration in a multi-ethnic school has been analyzed in [14] with the AHP method (Analytic Hierarchy Process) introduced by [18] and deepened in [19]. The teachers' point of view has been investigated, because they are considered privileged observers. The purpose of our research is to determine the critical variables; in other words, the variables that have the greatest effects on the degree of scholastic integration.

The complex general objective (GO = "The foreign student's scholastic integration") of analysis has been divided into two particular sub-objectives (A1 = Interpersonal communication; A2 = The degree of scholastic profit). With the AHP procedure to each one of these sub-objectives has been associated a weight with respect to the general objective.

Moreover, a set of variables has been determined which allow us to give an implicit definition of the general objective. The variables considered are: B1 = interaction and relations in class with peers; B2 = relations in the extra-curricular time; B3 = interaction and relations in class with teachers; B4 = relations with the belonging environment; B5 = linguistic-expressive abilities; B6 = logic-mathematical abilities; B7 = manual skills; B8 = group activities; B9 = sport skills.

In order to assign a measure of the degree of importance of the variables, a committee of 12 autochthonous teachers T(r), r = 1, 2, ..., 12, very expert and interested in the problem, has been formed. They has been requested to express their judgements with three matrices of pairwise comparisons: the matrix A(r) of the sub-objectives A1 and A2, w.r. to the general objective; the matrix M(r) of the variables Bi, w.r. to A1; the matrix N(r) of the variables Bi w.r. to A2. We emphasize that these teachers are not a sample of a statistical collective, but they are a group of decision makers chosen for their competence in the problem.

The criterion to compile the matrices is the one suggested by Saaty. Let X1, X2, ..., Xm be the objects to compare (e.g., sub-objectives or variables). Every teacher T(r), if considers Xi preferred or indifferent to Xj, then is requested to estimate the importance of Xi with respect to Xj using one of the following linguistic judgments: *indifference*, *weak preference*, *preference*, *strong preference*, *absolute preference*.

Then the linguistic values are expressed as numerical values following the Saaty fundamental scale: indifference = 1; weak preference = 3; preference = 5; strong preference = 7; absolute preference = 9. If the object Xi has one of the above numbers assigned to it when compared with object Xj, then Xj has the reciprocal value when compared with Xi. Then a pairwise comparisons matrix A = (a_{ij}) with m rows and m columns is associated to the m-tuple (X1, X2, ..., Xm), where a_{ij} is the number assigned to Xi when compared with Xj.

Let Y(r) = $(y_{ij}(r))$ denote the generic element of the set A(r), M(r), N(r). The synthesis of the teachers' opinions is made by considering, for every each pair (i, j) of indices, the geometric mean $G^Y = (G^Y_{ij})$ of the opinions of the teachers, defined by the formula: $G^Y_{ij} = (y_{ij}(1)y_{ij}(2)...y_{ij}(12))^{1/12}$.

The weights associated to the elements represented to the rows of G^Y are the components of the normalized positive eigenvector E^Y associated to the principal eigenvalue λ^Y of the matrix G^Y. Moreover, if n is the number of objects submitted to pairwise comparisons, Saaty suggests to consider the consistence index $\mu^Y = (\lambda^Y - n)/(n-1)$, and assumes the judgments are strong coherent if $\mu^Y \leq 0.1$.

As to the matrix G^A, of order 2, the consistence is evident, and $\lambda^A = 2$, $\mu^A = 0$, $E^A = (0.874, 0.126)$. For the matrix G^M and G^N we have:

$\lambda^M = 10.344$, $\mu^M = 0.168$, $\lambda^N = 10.306$, $\mu^N = 0.163$, with eigenvectors

$E^M = (0.303, 0.211, 0.164, 0.042, 0.090, 0.058, 0.039, 0.057, 0.031)$,

$E^N = (0.302, 0.107, 0.165, 0.055, 0.124, 0.104, 0.054, 0.059, 0.025)$.

The not strong coherence is justified by the differences of the opinions of the 12 teachers. Let $E^{M,N}$ the matrix having as rows E^M and E^N. The matrix product $W = E^A E^{M,N}$ is the vector of the scores of the variables. We obtain:

$W = (0.303, 0.198, 0.164, 0.044, 0.094, 0.064, 0.041, 0.057, 0.030)$.

By previous analysis we can underline that, in order to obtain the foreign students scholastic integration, in the opinion of the group of teachers, the most important variables are, in order of preference: B1 = interaction and relations in class with peers; B2 = relations in the extra-curricular time; B3 = interaction and relations in class with teachers.

3 Intercultural Integration from the Students' Point of View and Statistical Analysis of the Results

To better understand the phenomenon of intercultural integration in the school an empirical research has been carried out - through the administration of a questionnaire - among autochthonous and foreign students attending the second and third year of first degree secondary school in Abruzzo Region (Italy). 1314 students have been interviewed, of which 881 Italians, 317 foreigners and 116 children of mixed couples.

For the analysis of Italian and foreign students scholastic experience in a school undergoing transformation characterized by a plural experience three dimensions

have been detected: *scholastic success* (regular attendance; self-evaluation of scholastic performance; commitment to school work; formative aspirations); *school as a place where to build social relations* (extra-scholastic friendship; relation with school friends); *school as a place where to build citizenship and intercultural relations*.

To obtain the aims of the research the following variables have been considered: *citizenship* (Italians, foreigners, children of mixed couples); *gender*; *socio-cultural capital of the family*; for the foreigners also the *age of immigration* and *areas of provenance* have been considered.

Scholastic Success. A datum that definitely needs attention is the gap between the scholastic success of foreign and autochthonous students. On the whole the foreigners achieve lower scholastic results, engage themselves less in studying, have a lower promotion rate, abandon school and they enroll in professional schools with a higher attendance compared to autochthonous. The factor of the time of permanence in the welcoming country acts in favor of a better scholastic performance but, to support what has been detected in other researches on the theme, it is not a sufficient element to grant a performance level equal to the one of the autochthonous.

These data confirm what has been generally supported in literature concerning the fact that migration is a structural element able to determine difficulties in the course of the formative pathway, above all at an initial level [16].

>From the research it is evident that scholastic success is highly influenced by the different gender and by the socio-cultural capital of the family. Female students, both autochthonous and foreigners, have better scholastic pathways compared to boys their same age. Difference in gender show a significance but almost identical compared to the three groups (Italians, foreigners, children of mixed couples).

Above all students from families belonging to higher social capital and families with a medium-high educational background are the ones to achieve higher scholastic success, and they are mainly orientated at choosing a grammar school and to attend university [5, 7, 12].

School as a Place where to Build Social Relations. As regards the school as a place where to build social relations, a first evident result is that the majority of the pre-adolescents interviewed has made friends above all in the scholastic context. It is noticeable, furthermore, an internal differentiation to subgroups, because the percentage of Italians that has met friends above all at school is significantly higher than the foreign one.

Another important datum that is worth to notice is the one related to a priority or less frequentation of peers with the same citizenship, because more than half of the non Italian students goes out above all with boys and girls with the same nationality.

In the group of foreigners it is relevant the influence of the immigration age and macro area of provenance factors. Above all the students that came to our country as older children that meet their friends in an extrascholastic environment and particularly among the peers from the same national group.

Therefore, school has an important role in the socializing process, above all at the same time - as has already been highlighted in literature [5] - it is evident that not always the relations that are established within the scholastic social area are extended to outside the world.

School as a Place where to Build Citizenship. Interesting results regard also the school as the building of a multiple citizenship and of intercultural relations - as openness towards diversity and sharing common values.

From the analysis of the data it is evident that a strong orientation of the students towards the concept of the school as a place where to learn the history of all the different people and cultures (European and extra-European) instead of a Euro-centered one, and at the same time towards the idea of the necessity of teaching the Italian Constitution, that is to say of the fundamental principles that are at the basis of the social living together [13, 1].

The opinions of the preadolescents are more articulate and highly conditioned from the nationality factor referring to the topic of teaching religion in a public school. Substantially there is a difficulty to think that school can offer the opportunity to open the students' mind in a cross-religious key, above all among the students with a Chinese origin [4].

As regards gender differences, autochthonous and foreign girls are more inclined than boys towards a concept of school as a place of intercultural learning.

Besides, it is noticeable that a high cultural capital of the family has a positive role in the formation of the young generation of intercultural relations and of a multiple citizenship.

Conclusions

From the investigation of scholastic experience of Italian, foreign and children of mixed couples students it is evident that foreign students, even the ones born in Italy, have lower scholastic success than the Italians.

In line with other studies on the theme, it emerges also the transversality of the positive action that a higher socio-cultural capital of the family has an influence on the scholastic success of the children [5, 7, 12].

It seems that the Italian school still has to take further steps to give all the students equality in the educational pathways and to allow the "new generations" of immigrants to obtain human capital that, together with material resources and social capital, are the basis of dealing effectively with the regulations of the social context and to break the vicious circle of disadvantage.

School is confirmed to be as the main place of human capital production in the group of peers [8], furthermore it is evident that not always the relations built within the scholastic social space are extended outside the school environment.

Among the preadolescents intercultural relations start to develop, above all the results of the research show how the road is still at the beginning. In such a vision, the idea of intercultural education has an important role and it takes on the new form of education of the citizenship that includes a intercultural dimension and that has as objectives openness, the respect for diversity, equality, and sharing of common values and social cohesion [13].

Multi-religiosity, that is evident in scholastic institutions, can be exploited through a dialogic comparison, intended at favoring the development of attitudes like respect and reciprocal exploration of the self [4, 15].

References

1. Ambrosini, M. (2008). Un'altra globalizzazione. Bologna: Il Mulino.
2. Bauman, Z. (1998). Globalization. The Human Consequences. Cambridge-Oxford: Polity Press, Blackwell.
3. Beck, U. (2003). La società cosmopolita. Prospettive dell'epoca postnazionale. Bologna: Il Mulino.
4. Besozzi, E. (2008). Culture in gioco e modelli di integrazione nella scuola italiana. In M. Clementi (Eds.), La scuola e il dialogo interculturale. Quaderni Ismu, 2/2008 (pp. 25–38). Milano: Vita e Pensiero.
5. Bourdieu, P. (1980). Le capital social. Rôtissoires. Actes de la Recherche en Sciences Sociales, n. 3,31.
6. Cesareo, V. (2004). L'Altro. Identità, dialogo e conflitto nella società plurale. Milano: Vita e Pensiero.
7. Coleman, J., S. (1988). Social Capital in the Creation of Human Capital. American Journal of Sociology, vol. 94, 95–121.
8. Coleman, J.S. (1991). Foundations of Social Theory. Boulder: Westview Press.
9. Council of Europe (2008). White Paper on Intercultural Dialogue. Living Together s Equals in Dignity, Strasburgo: www.coe.int/dialogue.
10. Farley, R. & Alba, R. (2002). The new second generation in the United States. International Migration Rewiew, vol. 36, n. 3, 669–701.
11. Giovannini, G. (2008). La scuola. In Tredicesimo rapporto sulle migrazioni 2007 (pp. 131–143). Milano: Fondazione Ismu, Franco Angeli.
12. Kao, G. (2004). Parental influences on the educational outcomes of immigrant youth. International Migration Rewiew, n. 38, 2, 427–449.
13. Kymlicka, W. (1995). Multicultural citizenship. Oxford: Oxford University Press.
14. Maturo, A., & Contini R.M., (2009). Application of the Analytic Hierarchy Process to the Sociological Analysis. In Proceedings of ISAHP2009, (paper 47, 1–13). Pittsburg, Pennsylvania.
15. Mentasti, L., & Ottaviano, C. (2008). Cento cieli in classe. Milano: Unicopoli.
16. Portes, A. (1995). The economic sociology of immigration. New York: Russel Sage Foundation.
17. Portes, A., & Rumbaut, R.G. (2001). Legacies. The story of the migrant second generation, Berkeley-New York: University of California Press-Russel Sage Foundation.
18. Saaty, T. L. (1980). The Analytic Hierarchy Process. New York: McGraw-Hill.
19. Saaty, T.L.. & Peniwati, K. (2007). Group decision-making: Drawing out and reconciling differences. Pittsburgh: PA: RWS Publications.
20. Sassen, S. (2007). A sociology of globalizzazion. New York: Norton & Company.

Mastering Complexity of Social Work – Why Quantification by Moral Scales and by Inspiring Diagnosis of Biographical Factors may bring more Effectiveness and Humanity

Bo Hu, Markus Reimer, and Hans-Rolf Vetter

1 Problem Statement

The more modern democratic welfare societies are confronted with dynamically developing social and economic differentiation, the higher the requirements on decision making and other organizational processes in social services and social work. Besides limited financial resources and legacy organizational infrastructures, their identity, their prevalent structures of thinking and discourses, their professional individual requirements as well as their moral and socio-political legitimization as qualified "interest neutral" practice facing increasingly heterogeneous clients make it hard for the organizations in this area to meet these challenges.

In this paper we want to line out that the upcoming and overdue restructuring of social work has to substantiate its basic approach in a consistent strategy of quantification. The following questions thus are raised on different levels operating consequently a rationalized type of diagnosis:

- Which measurements have to be introduced to achieve a greater level of efficiency and a considerably higher plausibility of financial and professional support?
- How can the reproducibility and verifiability of decision making processes of social work be augmented?

Bo Hu
Universität der Bundeswehr München, 85577 Neubiberg, Germany
e-mail: bo.hu@unibw.de

Markus Reimer
Institut für Päd. Controlling und Soz. Qualität, D-93437 Furth/W., Germany
e-mail: m.reimer@quin-neustart.de

Hans-Rolf Vetter
Universität der Bundeswehr München, 85577 Neubiberg, Germany
e-mail: hans-rolf.vetter@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_95,

- How can professionality be assured against ideological and in most of the cases non-specialist interventions (e.g. irritating lay person knowledge)?
- How might a higher efficiency be achieved via specialization?
- In how far must the hitherto prevailing thinking and working habits be made the topic of sustained measures of re-professionalization with respect to the necessities of further professional and technical networks?
- To which extents has a prognosis of requirement, which is not only superficially registering, but already also highly differentiated in its approach of production, to be performed with respect to the growing heterogeneity of the population?

2 Minimization of Moral Dilemmas and Enhancement of Successes of Labor Market Re-Integration Using Biographical Factors-Tableau

Both individual case work and the ensuring of moral principles demonstrate that the implementation of quantification procedures is necessary in multiple respect. On the condition of a fundamental shortage of resources social workers intend (to be able) to maintain the provision of the population with differentiated social services, to organize appropriate and authentic aid and simultaneously to augment the professional quality of the supply. A re-thinking or a fundamental mental caesura in social work and the social services seems to us to be indispensable as an entry requirement for the establishing of more effective and more transparent procedures. Two arguments are crucial here in particular:

- Even if assumed – which is rather unlikely because of the shortage of resources that has increasingly become apparent – that the number of households inquiring social work will not increase a higher level of social heterogeneity will ensue from the tendency of individualization of the society. Not only the "anomy problems" will accumulate gravely, but also that social work which has emerged basically from the combating of the classical complex of problems of poverty and des-integration, will be confronted with historically new types of customers and clients, especially from the modern middle classes.
- Furthermore, the expansion of the modern knowledge-based society demands increasing enhancement and availability for use of the empirical examination methods and diagnosis techniques as well as a reciprocal penetration of the modern disciplines in the sense of interdisciplinary approach. This might be the one and only successful professional basis to guarantee increasingly reality adequate analyses, prognoses and forms of intervention of social work. Here the combination is pending, we are referring to as a central approach in this contribution; the mutual mediation of educational science with the forms of practice and systems of addressees of social work also on the organizing basis of information technology.

Social work and the modern forms of mediation and support management have to harmonize efficiency, budget, professional excellence and individualization with one another in appropriateness to the reality in the course of operationalized measures.

So let‘s have a look to two typical scenarios – moral based decisions for help and re-integration of jobless people into the labor market – to line out the advantages of a quantified model of decision making and fixing of adequate operation systems.

2.1 A Support-Priority-Value as an Aid to Decision-Making in Social Work?

As a basic principle, the forms of practice and the self-conception of the service providers are implying an infinite claim of moral legitimization. In a modern welfare society this claim is confronted with the reality that because of increasingly scarce budgets decisions have to be made about which needy persons in which problem situation have to experience social care and how they can be helped subjectively. The infinite moral claim clashes with the finite availability of mobilizable resources. This enforces a priorization of the resources invested within the framework of mandatory catalogizations. So that these will not sample out arbitrarily, is ensues logically, that the relevant concepts of aid, intervention and service work will have to be able to be structurally covered on the basis of differentiated quantification, to be categorized functionally and configured as reliably operationalized practice.

Thus, it is also the code digits that have to be ascertained that reflect the dimensions of the emerging social context of problems or of the individual case, respectively, and which simultaneously disclose via their numerics, to which extent, with which prospect of success and in which placing, resources can be made available. This takes place against the background, that *theoretically* and above all from a "moral" perspective, it would be possible always to help more, than this is possible *practically* on the basis of scarce resources. And this dilemma raises thus further questions: Might the amount allocated, with which a support is initiated for a person, perhaps in a considerably better way not be firstly invested for another form of support and secondly for another person? According to which criteria is a social worker meant to judge without a doubt, which person he or she should help and which person not, and which support should be initiated in which extent?

It becomes blatantly obvious here that the answers to these questions are linked extremely closely to the moral categories: Any decision, which is made *in favor* of a needy person, is simultaneously a decision *against* another needy person. And in the resolving of this moral dilemma lies a second essential advantage of the here preferred quantification strategy.

Thus, decisions in social work, which are structured systematically traceably and therefore also furthermore reliably, are more easily to be made both for the needy person as on the other hand also for the social worker. The allegation of an (actual or alleged) arbitrariness will be avoided to a large extent by the standards, which

signify at the same time an impulse of professionalization for the organization of social work and social services. As a consequence of the same standards and their logifization as an operation mode, the total volume of the resources will then be used more effectively and the necessary aspect of humanity is abetted. As a possibility of a solution, thus, also the implementation of a type of "support priority value (SPV)" is imaginable. The SPV might disclose then, where and how the existing resources can be allocated most effectively and most efficiency. The SPV might be composed as a product of the value of diagnosis (diagnosis validation), support value (potential of the peculiar organization to offer the support requested), and the result value (prognosticated positive result from individual and social level). The higher the SPV, the better the resources seems to be invested. This option as well seems to make clear once more that the relatively clear Win-Win situations might arise for the "producers" of social as well as for their addressees in the moment, in which the carefully-prepared quantification procedures are replacing the hitherto "accidental", at best experience-based decision-making dilemmas under the vigilance of quality requirements.

2.2 The Quantification of Case-Related Diagnostics by Means of the Biographical Factors Tableau (BFT)

Identical as mentioned under 2.1 is the quantified ascertainment of the subjective complex of problems, which are linked with the integration or re-integration, respectively, into the labor market. The answer in this case is: "biographical factors-tableau (BFT)". By support of this modeling approach which has started to be tested in an empirical manner we have sought to harmonize the pedagogical, informational, functional and economic demands within a special framework of a model of analysis and prognosis. Based upon the biographical foundations that a person exhibits with respect to employment, lifestyle, dwelling, mobility etc., a quantifiable categorization regarding further opportunities of arranging employment is conducted according to this model. The model exhibits two decisive reconstructions for the current practice of support:

- Beside the statistical and socio-structural facts the subjective assessments and evaluations towards one's own life situation have found their place in the model.
- For re-integration into the labor market, biographical, social and infrastructural parameters have a much wider influence upon the success of arranging employment than the immediate qualification structure of the working capacity. The working capacity is conditioned quite considerably by external-job-specific framework conditions and to a lesser extent by the direct fit of the qualification inquired by enterprises.

The results are thus made per factor – Ten such individual factors are concerned altogether – on a scale between 0 and 1. The value "0" signifies this person's absolute absence of opportunities of being able to be re-integrated enduringly into the labor

market within the subsequent three years, at a value of "1" – with respect to this factor – the opportunities of a direct, successful re-integration into the labor market amount to 100% per paper form.

Beyond the individual factors, however, the "case-typical" overall evaluation in the sense of a solid prognosis for the future is crucial. With an average value of below 0.5 a successful re-integration into the labor market is not provided without an enduring support of individual factors (e.g. mobility, housing, social environments) on the hitherto starting level. The closer the total value tends towards the direction of "1", the more dynamical the opportunities of re-integration are increasing realer.

It is intended both to impart a precise listing of and orientation towards their situation in life to the individual persons seeking employment, so that they will be in position to start a "methodical conduct of life" (see Weber 1956) by means of scientifically-substantiated analyses. Moreover, however, above all the actual practice of re-integration is intended to be de-bureaucratized and de-mystified, i.e. to be positioned upon rational, comprehensible and simultaneously scientifically-substantiated foundations of support. Our own manifold individual empirical studies show that precisely the previous practice of support suffers from making use in a one-sided manner of authoritarian methods, merely in perspective and provided with high levels of uncertainties of prognosis, in which both the real as well as the future support-capable potential of the clients and customers and the complex role of blocking factors in a subject's entire everyday-life-organization are merely covered inadequately and smoothed over by "ideological" assumptions.

3 What has to be Done – Toward Methodological and Logistical Costs of Quantification and Professionalization

In the final analysis, this catalogue of problems points thus to three milestones:

1. The quantification of social work changes its standards of work and its quality measurement to an extensive extent.
2. It will still be the matter of individualized services which are provided within an authentic relationship – inclusive of the co-producer issue, these will, however, than have to yield more than ever to the verdict of uniform norms within the framework of identically practiced regulations of procedures. This excludes arbitrariness and coincidence to a large extent, but certainly also a part of the tolerances, which have been due to the procedure moderation, which has been hitherto related to a larger extent to the direct person. We are dealing here certainly with a certain revocation of "individualization" as a moral concern of social work. The individual case becomes a type-shaping constellation of numerically recorded and controlled factorial attributes.
3. The third dimension affects social work deeply to its historic core: It touches its moral fundament: Because any decision which is made according to a catalogued and at the same time quantified system in favor of or against the aid

of a needy person is simultaneously also a decision in favor or against the aid for another needy person. Here a moral dilemma arises for the political self-conception of social work in so far, as the decision of "aid" is basically actually never doubted morally, whereas the decision for "no aid" is evaluated to be morally highly alarming, if not even condemnable. Here a further problem becomes apparent: Moral stands contra to predictability and efficiency. Because in the "ancient" system of social work, it is neither a matter of the quality of aid nor its professional supplementation, but much more only of the fact that aid is offered at all.

4 Conclusion and Outlooks

The approach, the making of decisions on the basis of quantification, is basically not new. In economy and administration, above all, however, also in humane service sectors as the medicine, this is the daily practice. In addition, also with respect to the multi-causalities implied, as it has to be assumed for the spheres of modern social work. To what we intended to point with an initial entry into the quantification issue of social services is, that success, effectiveness and humanity may well incur a satisfying alliance also in such spheres, which have been referred to on the basis of their immanent complexity and contingence conflicts (Willke 1996) as being little predestinated for operative formulations of the problem, which have been perceived basically on constricted conditions of individualization and effectiveness. What we sought to propagandize here is also that social work and social services can act under in the meantime similar conditions of model and empiricism, as for example the medicine and psychology and that is thus should get its suggestions, where positive experiences are available already in a manifold manners compared with these conditions.

References

1. M. Reimer. *Pädagogisches Controlling*. Baden-Baden, 2009.
2. W. Sinnot-Armstrong. *Moral Psychology (3 volumes)*. MIT Press, Cambridge, 2008.
3. H.-R. Vetter, Bo Hu, and N.-A. Bauer. *The Cultural Demands toward Successful Self-Organization – An Empirical Example of Individualized Programs of Advancement of Qualified Unemployed*. ECER 2010, Helsinki, 2010.
4. Max Weber. Wirtschaft und Gesellschaft im allgemeinen. In *Ders.: Soziologie, Analysen, Politik*. Stuttgart, 1956.
5. Helmut Willke. *Systemtheorie I: Grundlagen – Eine Einführung in die Grundlagen sozialer Systeme*, 2. Auflage. Stuttgart, 2006.

III.7 Young Scientist's Session: First Results of On-Going PhD-Projects

Chair: Prof. Dr. Stefan Pickl (Universität der Bundeswehr München)

The Young Scientist's Session encourages young researchers to present initial results of their PhD projects. There is no agenda or concentration on specific topics.

München 2018

In fact, classical and interdisciplinary OR topics are welcome, to include extraordinary subjects such as "Operations Research and Sports" (Supporting Munich's application for the 2018 Olympic Games, http://www.muenchen2018.org/bewerbungsthemen/schueler+studierende/index.html).

Oberwolfach style

The talks or poster presentations should be within 15–45 minutes in length, and the Session, which is offered for the first time at the International Operations Research Conference, will have a special "seminar" character similar to the Oberwolfach style. The scientific interaction, the exchange with the other PhD students and the presentation of new approaches, are the primary benefits of this session.

Award Paper

The papers by Gabrio Caimi, Benjamin Hiller, Debora Mahlke and Florian Sahling are GOR dissertation award papers, and the paper by Rainer Hoffmann is a GOR diploma thesis award paper.

Scheduling in the Context of Underground Mining

Marco Schulze and Jürgen Zimmermann

1 Introduction

During the last five decades, numerous publications[1] have appeared concerned with the application of optimization methods in the mining industry. Most of them focus on long-term production scheduling for underground mining, e.g. [7] as well as open pit mining, cf. [4]. In contrast, this paper addresses the short-term underground mine production scheduling problem that can be defined as specifying the sequence in which blocks should be removed from the mine. The aim is to minimize the makespan subject to a variety of constraints, because the management of the mining company is interested in an efficient utilization of the resources. The constraints relate to the mining extraction sequence, resource capacities and safety-related restrictions.

The extraction of the examined German potash mine is done by room-and-pillar mining. In this mining system the mined material is extracted across a horizontal plane while leaving pillars of untouched material to support the roof of the mine. Thus, open areas (rooms) emerge between the pillars. As mining advances, a grid-like pattern of rooms and pillars is formed. There are two types of room-and-pillar mining: conventional mining and continuous mining. Except for some special applications, the excavation of potash is based on the former type involving drilling and blasting. This kind of underground mining is characterized by the following consecutive sub-steps (operations), that can be defined as a production cycle:

1. scaling the roof
2. bolting the roof with expansion-shell bolts
3. drilling large diameter bore holes
4. removing the drilled material
5. drilling blast holes
6. filling the blast holes with an explosive substance
7. blasting
8. transportation of the broken material to a crusher.

Marco Schulze, Jürgen Zimmermann
Clausthal University of Technology, Institute of Management and Economics, Operations Research Group, e-mail: marco.schulze@tu-clausthal.de, juergen.zimmermann@tu-clausthal.de

[1] Good surveys are given by [6] and [8].

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 611
DOI 10.1007/978-3-642-20009-0_96,

For each processing step except for the blasting step one special mobile machine is required. In order to excavate one block[2] of a certain underground location (for example block 8 at location a2 in Fig. 1) it is necessary that all sub-steps of the preceding production cycle have been finished, e.g. the preceding block (block number 7 in Fig. 1).

After the completion of the first step of the production cycle, the remaining operations ought to be finished within a certain time limit τ. If an operation (except for the transportation step that is processed by a sheltered loader) cannot start within the time limit, a security precaution is needed in which the roof is scaled once more. Then, the next operation of the original cycle can be resumed (see Fig. 2, where τ has elapsed for a job after completion at stage 3).

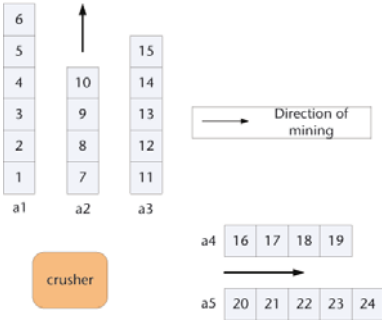


Fig. 1 Schematical view of a mining region that consists of five locations (a1-a5), several blocks per location (4-6) and one crusher. For a better orientation the blocks are consecutively numbered.

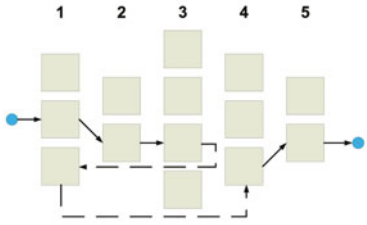


Fig. 2 Hybrid flow shop design of the excavation process. The dashed line symbolizes that the time period τ is exceeded. Consequently the job has to visit stage 1 (scaling the roof) once more.

The described excavation process represents a common manufacturing environment that can be identified as a hybrid flow shop scheduling problem[3]. The hybrid flow shop problem is a generalization of the classical flow shop problem. There are K production stages (i.e. sub-steps of the production cycle) in series, separated by unlimited intermediate buffers, and each stage k consists of $M^{(k)}$ unrelated parallel machines. The jobs (i.e. blocks) have to visit the stages in the same order starting from stage 1 through stage K. A machine can process at most one job at a time and a job can be processed by at most one machine at a time. Preemption of processing is not allowed. The scheduling problem consists of assigning jobs to machines at each stage and sequencing the jobs assigned to the same machine so that the makespan is minimized. In contrast to this "standard" form of the hybrid flow shop problem, cf. [1], specific restrictions associated with the underground mining production cycle have to be considered.

[2] Usual dimension: 12 meters wide, 6 meters high, 6.3 meters deep.

[3] More details about this problem and a comprehensive literature review are given in [5].

2 A Mixed Integer Linear Programming Formulation

We give a detailed mathematical description of the production cycle as a mixed integer linear programming problem. The used notifiers are gathered in Table 1.

Table 1 Notifiers

Sets and indices	
α, β	indices, $\alpha, \beta \in \{1,2\}$
k	production stage index, $k = 0,1,2,\ldots,K$
l	machine index, $l = 1,2,\ldots,M^{(k)}$
$\mathcal{N}^a$	set of jobs at location a, $\mathcal{N}^a = \left\{\left(\sum_{\eta=1}^{a-1} n_\eta\right)+1,\ldots,\sum_{\eta=1}^{a} n_\eta\right\} = \{\min\mathcal{N}^a,\ldots,\max\mathcal{N}^a\}, \forall a$
j, r	job indices, $j, r = 1,\ldots,n$
Parameters	
a	number of underground locations, $a = 1,\ldots,A$
$M^{(k)}$	number of parallel machines at stage k
n_a	number of jobs at location a
n	total number of jobs, $n = \sum_{a=1}^{A} n_a$
p_{jkl}	processing time of job j on machine l at stage k
τ	critical time period causing recirculation
Q	large number, $Q \geq \sum_j \sum_k M^{(k)} \cdot \max_{l \in M^{(k)}} \{p_{jkl}\}$
Variables	
C_{jk}	completion time of job j at stage k; C_{j0} symbolizes the earliest start at the first stage
C_{j1}^2	completion time of job j at the first stage, when job j is processed for the second time
$C_{\max}$	makespan, $C_{\max} = \max_{\{j=1,\ldots,n\}} \{C_{jK}\}$
D_{jk}	specifies the time-lag between C_{jk} $(k = 2,\ldots,K-2)$ and C_{j1}
X_{jrk}	1, if job j precedes job r at stage k $(k \geq 2)$; 0 otherwise
$X_{jr1}^{\alpha,\beta}$	1, if job j precedes job r at the first stage, where job j is scheduled for the first time $(\alpha = 1)$ or the second time $(\alpha = 2)$ and job r is scheduled for the first time $(\beta = 1)$ or the second time $(\beta = 2)$; 0 otherwise
Y_{jkl}	1, if job j at stage k $(k \geq 2)$ is scheduled on machine l; 0 otherwise
Y_{j1l}^{α}	1, if job j at the first stage is scheduled on machine l for the first time $(\alpha = 1)$ or second time $(\alpha = 2)$; 0 otherwise
Z_{jk}	1, if τ is exceeded for job j after completion at stage k $(k = 2,\ldots,K-2)$; 0 otherwise
Z'_{jk}	binary variable, that ensures at most one recirculation for any job

$$\text{Minimize } C_{max} \tag{1}$$

subject to

$$C_{max} \geq C_{jK} \qquad \forall j \tag{2}$$

$$\sum_{l=1}^{M^{(k)}} Y_{jkl} = 1 \qquad \forall (j,k), k \geq 2 \tag{3}$$

$$\sum_{l=1}^{M^{(1)}} Y_{j1l}^{1} = 1 \qquad \forall j \tag{4}$$

$$\sum_{l=1}^{M^{(1)}} Y_{j1l}^{2} = \sum_{k=1}^{K-2} Z_{jk} \qquad \forall j \tag{5}$$

$$C_{j1} - C_{j0} \geq \sum_{l=1}^{M^{(1)}} Y_{j1l}^{1} \cdot p_{j1l} \qquad \forall j \tag{6}$$

$$C_{j+1,0} = C_{jK} \qquad \forall j \in \mathcal{N}^a \setminus \max \mathcal{N}^a, \forall a \qquad (7)$$

$$C_{j0} = 0 \qquad \forall j = \min \mathcal{N}^a, \forall a \qquad (8)$$

$$C_{jk} - C_{j,k-1} \geq \sum_{l=1}^{M^{(k)}} Y_{jkl} \cdot p_{jkl} \qquad \forall (j,k), k \geq 2 \qquad (9)$$

$$D_{jk} = -C_{jk} + C_{j1} + \tau + Q \cdot Z'_{jk} \qquad \forall (j,k), k = 2, \ldots, K-2 \qquad (10)$$

$$D_{jk} \leq Q \cdot (1 - Z_{jk}) \qquad \forall (j,k), k = 2, \ldots, K-2 \qquad (11)$$

$$D_{jk} \geq -Q \cdot Z_{jk} \qquad \forall (j,k), k = 2, \ldots, K-2 \qquad (12)$$

$$Z'_{jk} = \sum_{\gamma=1}^{k-1} Z_{j\gamma} \qquad \forall (j,k), k = 3, \ldots, K-1 \qquad (13)$$

$$Z'_{jK} = Z'_{j,K-1} \qquad \forall j \qquad (14)$$

$$\sum_{k=1}^{K-2} Z_{jk} \leq 1 \qquad \forall j \qquad (15)$$

$$C^2_{j1} \geq C_{jk} + \sum_{l=1}^{M^{(1)}} Y^2_{j1l} \cdot p_{j1l} - Q \cdot (1 - Z_{jk}) \qquad \forall (j,k), k = 2, \ldots, K-2 \qquad (16)$$

$$C_{jk} \geq C^2_{j1} + \sum_{l=1}^{M^{(k)}} Y_{jkl} \cdot p_{jkl} - Q \cdot (1 - Z'_{jk}) \qquad \forall (j,k), k = 3, \ldots, K-1 \qquad (17)$$

$$Q\left(2 - Y_{jkl} - Y_{rkl} + X_{jrk}\right) + C_{jk} - C_{rk} \geq p_{jkl} \qquad \forall (j,k,l,r),\ j < r,\ k \geq 2 \qquad (18)$$

$$Q\left(3 - Y_{jkl} - Y_{rkl} - X_{jrk}\right) + C_{rk} - C_{jk} \geq p_{rkl} \qquad \forall (j,k,l,r),\ j < r,\ k \geq 2 \qquad (19)$$

"Avoid overlapping at the first stage" (20)

$$C_{jk}, D_{jk} \geq 0 \qquad \forall (j,k), k \geq 2 \qquad (21)$$

$$C_{j1}, C^2_{j1} \geq 0 \qquad \forall j \qquad (22)$$

$$Y_{jkl} \in \{0,1\} \qquad \forall (j,k,l), k \geq 2 \qquad (23)$$

$$Y^1_{j1l}, Y^2_{j1l} \in \{0,1\} \qquad \forall (j,l) \qquad (24)$$

$$X_{jrk} \in \{0,1\} \qquad \forall (j,r,k),\ j < r,\ k \geq 2 \qquad (25)$$

$$X^{1,1}_{jr1}, X^{1,2}_{jr1}, X^{2,1}_{jr1}, X^{2,2}_{jr1} \in \{0,1\} \qquad \forall (j,r),\ j < r \qquad (26)$$

$$Z_{jk} \in \{0,1\} \qquad \forall (j,k), k = 1, \ldots, K-2 \qquad (27)$$

$$Z'_{jk} \in \{0,1\} \qquad \forall (j,k), k \geq 2 \qquad (28)$$

The optimization criterion considered is the minimization of the makespan, i.e. the latest completion time of the last operation of the production cycles, cf. (1) and (2). Constraints (3)–(5) guarantee that all operations are assigned strictly to one machine at each stage. Constraint set (4) ensures the assignment at the first stage if one job has to be processed for the first time and constraint set (5) ensures an assignment in case of recirculation[4]. The set of constraints (6) restricts the starting time of the first operation to be greater or equal to its release time from the fictitious stage 0. Eq. (7) and (8) determine the earliest start of processing at the first stage, distinguished between the first job of a location and the proceeding jobs.[5]

[4] If jobs may visit each stage several times, it is called re-entry or recirculation, c.f. [3] or [2].

[5] Eq. (7), (8) and Fig. 1 emphasize the chain precedence constraints for the jobs belonging to a certain underground location.

Constraint set (9) assures that a job cannot start its processing at stage k $(k \geq 2)$ before it is finished at stage $k-1$. Constraints (10)–(12) force $Z_{j,k} = 1$ if the time limit τ is exceeded after processing job j at stage k.[6] The aspect that the additional operation in case of recirculation of a job may occur at most once, is ensured by constraint sets (13), (14) and (15). The determination of a lower bound of C^2_{j1} for any job j that is processed twice at the first stage, is described by constraint set (16). The completion time of the next operation of the original cycle in case of recirculation is determined by constraint set (17). Constraint sets (18)–(20) prevent any two operations from overlapping in a common machine. The constraints (20) are respresentatives for a set of 8 constraints that avoid overlapping at the first stage, distinguished between four different cases: the competing jobs j and r visit the stage for the first time, only one job (j or r) has to be processed for the second time or both jobs visit the stage for the second time. Constraint sets (21)–(28) define the domains of the decision variables.

3 Computational Results

The computational results[7] for the model as introduced in Sec. 2 are given in Table 2. We generated fifteen scenarios, each with ten instances, where the number of parallel machines at each stage ($K = 7$ for all scenarios) and the total number of jobs were varied. Thus, the specification (2-2-1-3-2-3-1) in the first column symbolizes the number of parallel machines at each stage. The entries presented in brackets in the remaining columns describe the number of underground locations (the number of jobs at each location ranges from 1 to 8).

When solving the small instances (*15, 20 and 25 jobs*) we set the time limit to 1,800 seconds and the maximum allowed gap to 0%. The results for these scenarios show something quite interesting: in case where we always have two parallel machines, the computation times are always lower compared to the cases with only one machine or the configuration with one to three parallel machines. After analyzing the results more in detail, we could see, that the LP bounds were always better when solving the scenarios with two parallel machines and that seems to be the reason for the faster results.

All small instances could be solved to optimality in acceptable computation time, but to fulfill practical needs we have to consider larger instances with around 15-20 underground locations and approximately 120 jobs that have to be scheduled for an usual mining district. Consequently, we analyzed near practical instances with *60* and *100 jobs* each. For these scenarios we set the time limit to 200,000 seconds and the maximum allowed gap to 5%. When solving the configurations with only one machine and one to three parallel machines at each stage, we implemented lower bounds that considered the bottleneck stage (results are identified in boldface). The results show that our lower bounds lead to significant faster results compared to

[6] To this end, the auxiliary variable D_{jk} checks if the time limit τ is violated for the first time.

[7] We have used Gurobi 3.01 on a Windows PC (Intel, 2x3.2 GHz, 2 GB RAM).

the lower bounds that were generated by the solver itself in case of two parallel machines.

Table 2 CPU time for the scenarios in seconds

	15 jobs (5)*	20 jobs (5)*	25 jobs (5)*	60 jobs (10)**	100 jobs (15)**
1 machine per stage	1.16	6.82	40.14	**430**	**2,597**
2 machines per stage	0.75	2.20	6.89	3,204	$>$ 200,000
1-3 machines per stage (2-2-1-3-2-3-1)	1.32	10.30	53.08	**190**	**1,015**

* Gap: 0%, time limit: 1,800 seconds. ** Gap: 5%, time limit: 200,000 seconds.

4 Conclusions and Outlook

In this paper we investigated the problem of scheduling in the context of underground mining. We identified the problem as a hybrid flow shop problem and proposed a MILP formulation of this scheduling case with specific mining restrictions. First computational results show that our approach can be applied to small and near practical instances.

Within future research we will generate and implement more sophisticated lower bounds, especially for the case with two parallel machines. In addition, we will extend the model considering stochastic processing times and machine breakdowns. Moreover, it is necessary to develop heuristics in order to get good solutions for practical applications with $>$ 120 jobs more faster.

References

1. S. A. Brah. *Scheduling in a flow shop with multiple processors*. PhD thesis, University of Houston, 1988.
2. H.-W. Kim and D.-H. Lee. Heuristic algorithms for re-entrant hybrid flow shop scheduling with unrelated parallel machines. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 223(4): 433–442, 2009.
3. M. L. Pinedo. *Scheduling: Theory, Algorithms, and Systems*. Springer, New York, 3 edition, 2008.
4. S. Ramazan and R. Dimitrakopoulos. Recent applications of operations research and efficient MIP formulations in open pit mining. *Transactions of Society for Mining, Metallurgy, and Exploration*, 316: 73–78, 2004.
5. R. Ruiz and J. A. Vázquez-Rodríguez. The hybrid flow shop scheduling problem. *European Journal of Operational Research*, 205(1): 1–18, 2010.
6. E. Topuz and C. Duan. A survey of operations research applications in the mining industry. *CIM Bulletin: the Canadian mining and metallurgical bulletin*, 82(925): 48–50, 1989.
7. L. P. Trout. Underground mine production scheduling using mixed integer programming. In Australasian Institute of Mining and Metallurgy, editor, *Application of computers and operations research in the minerals industries, Brisbane, Australia, 9–14 July 1995*, volume 95,4, pages 395–400, Carlton, 1995. Australasian Institute of Mining and Metallurgy.
8. A. Weintraub, C. Romero, T. Bjørndal, and R. Epstein. *Handbook of operations research in natural resources*, volume 99 of *International series in operations research & management science*. Springer, New York, 2007.

Multi-Attribute Choice Model: An Application of the Generalized Nested Logit Model at the Stock-Keeping Unit Level

Kei Takahashi

Abstract This paper proposes an application of the generalized nested logit (GNL) model which is used in transportation science for product choice problems at the stock-keeping unit level. I explain two alternative nesting rules: *attribute separation* and *latent-class separation based on taste heterogeneity*. First, using the former nesting rule, I demonstrate that the GNL model is superior to the multinominal logit and the nested logit models in terms of reproducibility of choice probabilities. Second, using latter nesting rule, I reveal that the compromise effect, which is inconsistent with utility maximization, occurs in the GNL model, which belongs to the general extreme value family. This shows that the compromise effect is, in fact, consistent with utility maximization in random utility circumstances.

1 Introduction

In this paper, I propose an application of the generalized nested logit (GNL) model to a product choice model at the stock-keeping unit (SKU) level. Usually, the nested logit (NL) model is used for expressing similarity (i.e., heterogeneity) among alternatives of the discrete choice model [2], which relaxes the assumption of independence from irrelevant alternatives and considers the similarity among alternatives within the same nest. Such heterogeneity among brands or products is defined as *structural heterogeneity* in Kamakura et al. [1]. However, the NL model can formulate the structure nested by only one attribute; this implies that it cannot express complex heterogeneity among multi-attribute products. This can be accomplished by the cross-nested logit (CNL) and the GNL models. These models can be considered as extensions of the NL model.

Kei Takahashi
Department of Industrial & Management Systems Engineering, Graduate School of Creative Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.
e-mail: takahashi@toki.waseda.jp

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_97,

They are employed for expressing complicated relationships among alternatives of route or mode choice problems in the field of transportation planning. By applying the GNL model for the product choice model at the SKU level, it is possible to express systematically complex heterogeneity among multi-attribute products, which cannot be expressed using the NL model. Furthermore, I conduct an empirical analysis with scanner panel data of actual product lines (cola in plastic bottles). In this analysis, I demonstrate the superiority of the GNL model to the multinominal logit (MNL) model and the NL model in actual reproduction.

Finally, I prove that the GNL model can express the compromise effect which is inconsistent with utility maximization. Since the GNL model belongs to the general extreme value (GEV) family (see Wen and Koppelman [8]), it is possible that the compromise effect occurs under utility maximization.

2 The GNL Model

In the GNL model, let P_k represent the probability of alternative k; this value is given by

$$P_k = \sum_j P_j P_{k|j},\ P_j = \frac{\left(\sum\limits_{k' \in N_j} \left(\gamma_{k'j} \mathrm{e}^{V_{k'}}\right)^{1/\mu_j}\right)^{\mu_j}}{\sum\limits_j \left(\sum\limits_{k' \in N_j} \left(\gamma_{k'j} \mathrm{e}^{V_{k'}}\right)^{1/\mu_j}\right)^{\mu_j}},\ P_{k|j} = \frac{\left(\gamma_{kj} \mathrm{e}^{V_k}\right)^{1/\mu_j}}{\sum\limits_{k' \in N_j} \left(\gamma_{k'j} \mathrm{e}^{V_{k'}}\right)^{1/\mu_j}}, \quad (1)$$

where $P_{k|j}$ denotes a conditional choice probability of alternative k if nest j is selected, V_k capitalizes the definite utility of alternative k, N_j is an alternative set which belongs to nest j, and allocation parameter γ_{kj} is interrupted by

$$\sum_k \gamma_{kj} = 1\ \forall j,\ 0 < \gamma_{kj} \leq 1\ \forall k, j. \quad (2)$$

The similarity parameter μ_j is a measure of the degree of independence among the alternatives in nest j, and must satisfy $0 < \mu_j \leq 1\ \forall j$ for consistency with utility maximization. The similarity parameters are constant in the CNL model: $\mu_j = \mu\ \forall j$.

Theorem 1 *The GNL model belongs to the GEV Family and is consistent with utility maximization.*

Proof. See Wen and Koppelman [8]. □

The above equations do not show how to nest alternatives; they only focus on the structures. The CNL and GNL models are used for transport mode choice [6] and route choice [7] in actual problems. Vovsha and Beckhor [7] construct a structure which makes link-based nesting in the upper level and chooses route alternatives in the lower level. I propose similar nesting rules in sections 2 and 3 respectively.

Table 1 Attributes of Products (Cola in Plastic Bottles)

k	Brand	Volume	Sugar-free	Purchase opportunities
1	Coca-Cola	1,500ml		1,605
2	Coca-Cola	500ml		506
3	Coca-Cola	1,500ml	sugar-free	707
4	Coca-Cola	500ml	sugar-free	180
5	Pepsi	1,500ml		1,024
6	Pepsi	500ml		525
7	Pepsi	1,500ml	sugar-free	611
8	Pepsi	500ml	sugar-free	111

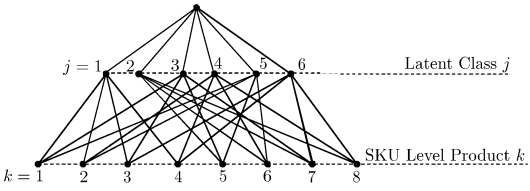


Fig. 1 Generalized nested structure of cola plastic bottles

3 Application

In this chapter, I apply the GNL model to an actual product choice problem. The data analysed here comprises one year of retail scanner data for cola in plastic bottles in two Japanese stores. The data includes 5,585 choices of 903 households. All product attributes and numbers of purchase opportunities are listed in Table 1.

Now, I propose a nesting rule for product choice. The products have three attributes — brand, volume, and whether they are sugar-free or not — and each attribute has two values. Consumers choose a product according to these attribute values; when nesting by attribute values, each nest represents a specific consumer taste. This paper calls this nesting rule *attribute allocation*. For example, consumers who prefer the brand 'Coca-Cola' chose the Coca-Cola nest at first and then chose one among the alternatives belonging to the Coca-Cola brand. In other words, this shows structural similarity among products which the NL model cannot capture sufficiently. The NL model can capture similarity in only one attribute; however, the GNL model can capture similarity in a voluntary number of attributes.

The definite utility function is specified as follows:

$$V_k = \alpha_1 X_{1k} + \alpha_2 X_{2k} + \alpha_3 X_{3k} + \alpha_4 X_{4k} + \alpha_5 X_{5k}, \tag{3}$$

where X_{1k} denotes the price of k, X_{2k} represents the brand dummy (Coca-Cola = 1, Pepsi= 0), X_{3k} indicates the volume of k, X_{4k} means the sugar-free dummy (sugar-free = 1, otherwise = 0), X_{5k} is a dummy variable indicating whether the product was purchased on the last purchase occasion (purchase on the last occasion = 1, otherwise = 0), and α_i is a parameter for the i th variable. The structure of cola plastic bottles in accordance with the above nesting rule is shown in Fig. 1.

Table 2 Comparison of Statistics

	[MNL]	[NL1]	[NL2]	[NL3]	[GNL]
LL	$-9,097$	$-9,071$	$-9,074$	$-8,858$	$-8,665$
AIC	18,205	18,155	18,163	17,729	17,380
$\bar{\rho}^2$	0.1695	0.1719	0.1716	0.1913	0.2082
Hit ratio	0.3993	0.4010	0.4005	0.4128	0.4147

Table 3 Estimated Parameters in the GNL Model

α_1	-0.0134^{**}	γ_{11}	0.2724^{*}	γ_{13}	0.2714^{*}	γ_{15}	0.4562^{**}
α_2	0.4847^{**}	γ_{21}	0.6083^{**}	γ_{33}	0.5433^{**}	γ_{25}	0.1700^{**}
α_3	0.0020^{**}	γ_{31}	0.0900^{*}	γ_{53}	0.6799^{**}	γ_{55}	0.2171^{**}
α_4	-0.3705^{**}	γ_{41}	0.2387^{**}	γ_{73}	0.9103^{**}	γ_{65}	0.0000
α_5	0.7309^{**}	γ_{52}	0.1030^{*}	γ_{24}	0.2216^{*}	γ_{36}	0.3667^{**}
η_1	0.1845^{*}	γ_{62}	1.0000^{**}	γ_{44}	0.1699^{*}	γ_{46}	0.5914^{**}
η_3	0.3423^{**}	γ_{72}	0.0897^{*}	γ_{64}	0.0000	γ_{76}	0.0000
η_4	0.5731^{**}	γ_{82}	0.0000	γ_{84}	0.2179^{*}	γ_{86}	0.7821^{**}
η_5	0.4625^{**}						
η_6	0.3974^{**}						

$*$ and $**$ indicate 95% and 99% significance in the t-test respectively.
η_2 is fixed at 1.00 .

To demonstrate the superiority of the GNL model, I compare it with the MNL model and the NL models nested by brands, volume, and whether sugar-free or not, which are named as NL1, NL2, and NL3 respectively. The maximum likelihood method is adopted for parameter estimation in each model. To avoid local maxima convergence, I estimate from ten different initial value sets.

Table 2 demonstrates the statistical superiority of the GNL model over other models. All statistics of the GNL model surpass those of other models. In particular, it shows that considering structural heterogeneity is a meaningful great contribution for the reproduction of product choice probabilities in the SKU level.

Table 3 shows the estimated parameters of the GNL model. Almost all of the parameters are 99% significant in the t-test. All similarity parameters are equal or lower than 1.00 and 99% significant in the t-test; therefore it seems that the alternatives in each nest are correlated.

4 Occurrence of the Compromise Effect in the GNL Model

The compromise effect can be explained as follows: When two alternatives A and B exist in two or more attribute spaces and neither dominates the other, and a new alternative C is added, it is relatively more attractive than A and B [4]. The strength of the compromise effect δ is given by

Table 4 Example parameters which lead to the compromise effect

DE_A	3.0	PQ_C	2.0	μ_3	0.9	γ_{B2}	0.6
DE_B	1.0	α_1	1.0	γ_{A1}	0.7	γ_{B3}	0.2
DE_C	2.0	α_2	1.0	γ_{A2}	0.2	γ_{C1}	0.1
PQ_A	1.0	μ_1	0.9	γ_{A3}	0.1	γ_{C2}	0.2
PQ_B	3.0	μ_2	0.7	γ_{B1}	0.2	γ_{C3}	0.7

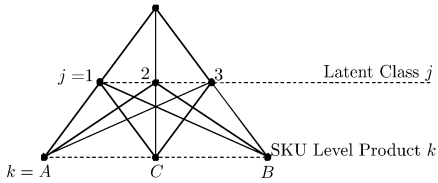


Fig. 2 Generalized nested structure leading to the compromise effect

$$\delta_A \equiv \frac{P_{C|\{A,B,C\}}}{P_{C|\{A,B,C\}}+P_{A|\{A,B,C\}}} - P_{C|\{A,C\}} > 0, \tag{4}$$

$$\delta_B \equiv \frac{P_{C|\{A,B,C\}}}{P_{C|\{A,B,C\}}+P_{B|\{A,B,C\}}} - P_{C|\{B,C\}} > 0, \tag{5}$$

in [5]. If only one side is positive, it is called polarization.

Four discrete choice models which can express the compromise effect are proposed in Kivertz et al. [3]. Kivertz et al. don't state relation to utility maximization. Furthermore, These models are not consistent with utility maximization with no constraints such as non-negativity of partial utility.

This paper thereby discloses that the compromise effect occurs in the GNL model which is consistent with utility maximization. The structure of the GNL model is assumed to include three nests and three alternatives, as shown in Fig. 2. The structure shown in Fig. 2 is like the one in the latent class logit model which has three alternatives and three latent classes. Then, I call this nesting rule *latent-class separation based on taste heterogeneity*. The difference from ordinary latent class logit model is in explanatory variables of segment size. I use logsum or category value as a explanatory variable of segment size with the GNL model instead of constants or demographic variables.

Suppose that the alternatives are cars which have two attributes: fuel efficiency DE and horsepower PQ. Therefore, definite utility function is assumed as

$$V_k = \alpha_1 DE_k + \alpha_2 PQ_k. \tag{6}$$

δ in this example can be calculated under the parameters shown in Table 4 as

$$\delta_A = \delta_B = \frac{0.336}{0.336+0.332} - 0.5 > 0. \tag{7}$$

Note that the compromise effect then occurs in the GNL model which is consistent with utility maximization in random utility circumstances.

First, what needs to be emphasized in this example is that $\gamma_{B1} > \gamma_{C1}$ and $\gamma_{A3} > \gamma_{C3}$. If these condition are violated, the compromise effect is never occurred in the GNL model. Second, it is important to note $V_A = V_B = V_C$. Kivertz et al. modify the definite utility leading to violation of utility maximization, however I modify the random utility by introducing allocation parameters and similarity parameters with the GNL model.

5 Conclusion

This paper has proposed an application of the GNL model for product choice problems at the SKU level. The GNL model potentially allows more flexible correlation structures than the hierarchical logit model. However, the application of the GNL model is limited to transportation problems, such as transport mode or route choices.

The first contribution of this paper is that it proposes a nesting rule called attribute separation for product choice problems at the SKU level. Indeed, this rule is one of the original contributions of this paper. For concrete results, I apply the GNL model to a choice problem involving cola in plastic bottles by utilizing the latent class nesting rule, and I then demonstrate the superiority of the GNL model over the MNL and NL models.

Furthermore, this paper makes a second contribution by proving that the compromise effect, which is inconsistent with utility maximization, occurs in the GNL model using latent class separation based on taste heterogeneity. Thus, it is shown that since the GNL model belongs to the GEV family, the compromise effect is consistent with utility maximization in random utility circumstances.

References

1. W. A. Kamakura, B. Kim, and J. Lee. Modeling Preference and Structural Heterogeneity in Consumer Choice. *Marketing Science*, 15(2): 152–172, 1996.
2. P. K. Kannan and G. P. Wright. Testing Structured Markets: A Nested Logit Approach. *Marketing Science*, 10(1): 58–82, 1991.
3. R. Kivertz, O. Netzer, and V. Srinivasan. Alternative Models for Capturing the Compromise Effect. *Journal of Marketing Research*, 41(3): 237–257, 2004.
4. I. Simonson. Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, 16(2): 158–174, 1989.
5. A. Tversky and I. Simonson. Context-Dependent Preference. *Management Science*, 39(10): 1179–1189, 1993.
6. P. Vovsha. Application of Cross-Nested Logit Model to Mode Choice in Tel Aviv, Israel, Metropolitan Area. *Transportation Research Record*, 1607(1): 6–15, 1997.
7. P. Vovsha and S. Bekhor. Link-Nested Logit Model of Route Choice: Overcoming Route Overlapping Problem. *Transportation Research Record*, 1645(1): 133–142, 1998.
8. C. Wen and F.S. Koppelman. The Generalized Nested Logit Model. *Transportation Research Part B*, 35(7): 627–641, 2001.

Judgment Based Analysis via an Interactive Learning Software for Modern Operations Research Education and Training

Arnold Dupuy, Victor Ishutkin, Stefan Pickl, and Romana Tschiedel

Abstract Distinct from traditional approaches and study modes, progress in information technology enables new techniques and procedures in educational processes. Focusing on independent or individual work, a judgment based component might be integrated in the traditional "teacher-pupil" education process. In this paper, a new Operations Research (OR)/Operational Analysis (OA) education software is presented, which is characterized by a strong embedded interactive component. This approach can lead to a special judgment based training method, and the classical Tactical Numerical Deterministic Model (TNDM), an empirically based combat simulation model, requires such special training techniques. We conclude with possibilities for such an extension of this famous program especially in the context of judgment based decision support processes.

1 Introduction

Judgment based analyses have become increasingly important for OR/OA-activities. The challenge is how decision situations might be trained, simulated and combined with new multimedia technologies. In a first step the advantages of such an approach

Arnold Dupuy
Virginia Polytechnic Institute and State University National Capitol Region Campus, Alexandria, VA, e-mail: Arnold.Dupuy@anser.org

Victor Ishutkin
Mary State University, Lenin sqr., Yoshkar-Ola, 424001, Russia, e-mail: izhutkin@yandex.ru

Stefan Pickl
Fakultät für Informatik, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg, e-mail: stefan.pickl@unibw.de

Romana Tschiedel
Fakultät für Informatik, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg, e-mail: romana.tschiedel@unibw.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_98,

will be described, while the second part, standard OR/OA aspects are addressed. Both sections will present the basis for a further development of the TNDM. We describe these requirements and conclude with a summary of the role of judgment based analyses within further decision support systems.

2 Individualised Training

The advantage of individualized training is each student is guided through the learning process at a speed which is adapted to his own learning progress. Within such processes, routine calculations and repetitions should be avoided. It is the individualized decision support activity which is the main center of interest! Therefore, how can such an approach can be embedded into classical OR/OA curricula? Is it possible to embed this process with a visualization component? Can we estimate the spatial experience of the students? Which new quality will define the scenarios? We will present a comprehensive approach to these questions based on four additional tasks:

- An intelligent structured process
- An interactive navigation software
- Scenario based approaches
- Strategic planning procedures supporting judgment based approaches

3 Training Functions to Support a Judgement Based Approach

If we develop a judgment based approach we integrate these four aspects within a holistic pedagogical "control" process. This is only possible if we realize a knowledge representation allows a quantitative and qualitative learning progress which consists of adaptive elements. Therefore, we have to structure the material/curriculum in order to "master the complexity" of a specific scientific area. For each pedagogical scenario we summarize and focus on:

- Problem statement
- Key idea
- Underlying methods/ procedures (algorithm, examples, exercises, test exercise)
- Planned errors
- Comparative analysis and examinations

These approaches should be supported by an interactive *navigation interface*. The key element is the system enables the user to understand the *holistic scientific* process. Examples and graphic illustrations underline the interactive process. In the following we might give an overview considering a few examples like "finding a maximum", "introduction to the simplex method", "matrix games" and "Hurwitz criterion". We are not able to go into detail in this paper, but the screenshots (see

Figs. 1 – 3) gives an impression of the advantages and characteristics of the software. The authors develop an integrated software package which contains at the moment a "classical optimization theory", "bargaining processes", "game theory" and "dynamical systems". These areas are essential for a judgment based decision making process. In the examples, training functions are used for judgment and information

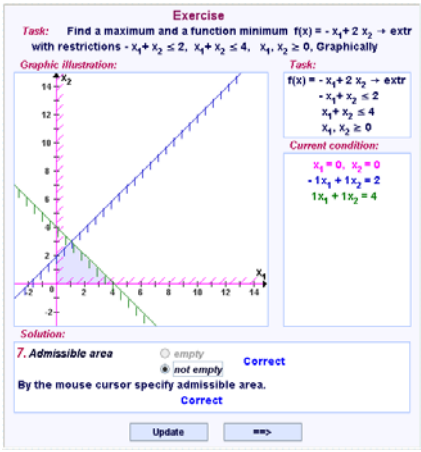


Fig. 1 Find a maximum – feasible areas are indicated. Interactive component – mouse cursor specifies feasible areas.

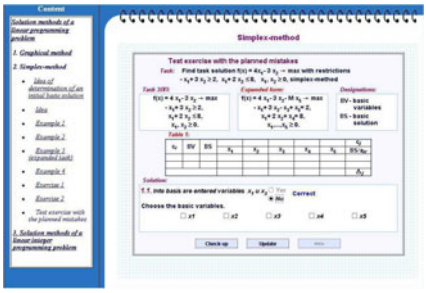


Fig. 2 Introduction to the Simplex algorithm.

enhancements which acquaint the user; they are inseparably linked to the comments which provide personalized feedback (see comments on the right side). Supervising functions are additionally applied – the degree is estimated by considering the errors, knowledge, level of understanding, etc.

Focusing on special (numerical) algorithms and decision procedures, it is a challenge to determine a correct selection and division of the teaching material into smaller parts. How can frequent control of knowledge be integrated, and which transitions might follow after certain phases? As we divide the learning process into a multi-stage process we prepare the basis for a so-called "multi-stage decision process" within special scenarios. These scenarios are a key element of the TNDM,

Fig. 3 Game theoretic model with min/max formulation.

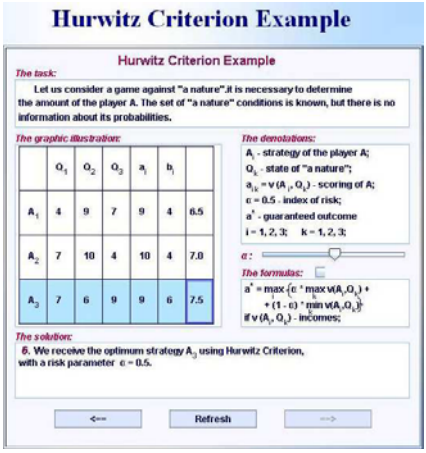


Fig. 4 Hurwitz Criterion – link to strategic planning processes.

which provide the user with planning and operational tasks and a variety of key decision points. The next chapter describes how we develop an education structure and asks how can we accurately divide the scenarios which are used in the software?

4 Training Scenarios

The system is characterized by relevant examples and suitable procedures. Different scenarios are embedded, which are characterized by judgment based components, while planned errors are used to confront the user with decision phases. The pedagogic element can be integrated into simulation software like the TNDM, an empirically based model used to adjudicate advance rates and combat losses in modern

warfare. The authors are working on an extension of the software via a suitable and comfortable graphical user interface.

5 Graphical User Interface

Key elements of the visualization tool are the combination of analytic models and graphical illustration techniques. Both refer to cognitive behavioral representations which enable a dynamic learning process. The process is characterized by the fact that:

1. Occurring changes might be stored: Therefore, the user can analyze how one aspect influence others
2. The user can directly influence geometric figures or ranges of functions
3. Changes of geometric objects can be tracked
4. Geometric objects on the computer screen can be changed smoothly, achieving performance of those or other conditions, and object characteristics can be analyzed

6 Tactical Numerical Deterministic Model (TNDM)

Our approach is characterized by a special individualized attempt: An example for a further application is the TNDM, an excellent example of a modern judgment based decision approach. We summarize its features and demonstrate how we will expand the software on the basis of a comfortable graphical user interface and an individualized training component.

The TNDM is an empirically based combat model with a database derived from historical research. It was developed by Colonel Trevor N. Dupuy, (USA, Ret.), from his concept, the Quantified Judgment Method of Analysis (QJMA), as presented in his two books, *Numbers, Predictions and War (1979)*[1] and *Understanding War: History and Theory of Combat (1987)*[2]. The QJMA has two elements:

1. Determination of quantified combat outcome trends based upon modern historical combat in more than 200 examples of 20th Century warfare, mostly World War II and the 1967 and 1973 Arab-Israeli Wars, and
2. Extrapolation of historical trends to contemporary and future combat on the basis of developments and changes in firepower and mobility technology.

The TNDM attrition methodology provides results consistent with those which occurred in historical engagements. By being historically based, the methodology is more scientifically justified than any methodology not consistent with historical experience. The TNDM is a proven, quick-reaction, inexpensive, computer-assisted mathematical simulation of air-land combat. It is suitable for planning, for analysis, and for examining a variety of situations, ranging from small-unit, low-intensity

combat action, to multi-day corps or army conventional battles. The software allows the historically-validated TNDM attrition methodology to be applied to a variety of analysis requirements, and as a combat "underlay" for designing and/or evaluating new combat systems/technologies. Before the outbreak of first Gulf War, analysts using the TNDM forecast US casualty rates far lower than predictions by any other model. In post-war assessments, adjustment of inputs to those actually experienced in the ground war, yielded TNDM attrition results within 5% of the historical experience of US forces during the period February 24-28, 1991.

There are two principal reasons for this. The patterns of TNDM attrition and advance rates in replication of historical combat closely correlate to those of historical experience over six decades, during which weapons and mobility technologies underwent sweeping changes. Furthermore, it is demonstrable that suppression effects of weapons on the battlefield are more important in battle outcomes than are attrition effects.

The TNDM can replicate the results of historical battles. Thus in historical and tactical instruction, it can demonstrate the results of "what ifs?", possible alternative outcomes if changes are made in basic inputs or in tactical decisions (see [3]).

7 Summary

The aim of this paper is an integration of a judgment based decision process in a classic "teacher-pupil" education environment. In the first part we present a new Operations Research (OR)/Operational Analysis (OA) education software. It is characterized by a strong embedded interactive component, and this interactive approach might lead to a special judgment based training method, using the TNDM for such special training techniques. The TNDM is described and joint proposals for such an extension are identified. The topic of the conference is "mastering complexity" and it is through such innovative pedagogical approaches that new methods for mastering complexity can be gained.

References

1. Trevor N. Dupuy. *Numbers, Predictions and War.* 1979.
2. Trevor N. Dupuy. *Understanding War: History and Theory of Combat.* 1987.
3. The Dupuy Institute. The tactical, numerical, deterministic model. http://www.dupuyinstitute.org/tndm.htm. Last accessed on 21/10/2010.

Non-Linear Offline Time Synchronization

Li Luo and Björn Scheuermann

Abstract In the design process of communication protocols it is necessary to perform repeated network communication experiments. Each run results in large event logs. The analysis of these logs is crucial to find and to understand unexpected behaviors and design flaws. Intrinsic to network communication these logs suffer from random delays, drop outs, and deviating clocks, which complicate the analysis.

Online synchronization protocols may interfere an experiment gravely and are unable to handle delays and drop outs. Offline synchronization approaches based on affine linear clocks using maximum likelihood estimation and least squares estimation are introduced by [3] and [2], respectively. We show that their approaches can be extended to non-linear clocks. The problem leads to a sparse linear program with a well-known structure, which can be readily solved by the interior point method. Under weak assumptions a consistency result is available for the least squares estimation.

1 Introduction

A real-world experiment with a computer network typically results in a set of event logs, one from each involved computer. For each event a log entry records a timestamp along with other information like, e. g., details on the transmitted or received data. Due to limited network connectivity and transmission errors these logs are inherently incomplete: not each computer will record each transmission. Thus, multiple logs are necessary to reconstruct the whole experiment. Yet timestamps, based

Li Luo
Chair of Mathematical Optimization, Heinrich-Heine-Universität, 40225, Düsseldorf, Germany; e-mail: luo@opt.uni-duesseldorf.de

Björn Scheuermann
Mobile and Decentralized Networks Group, Heinrich-Heine-Universität, 40225, Düsseldorf, Germany; e-mail: scheuermann@cs.uni-duesseldorf.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, DOI 10.1007/978-3-642-20009-0_99,

on local clocks in each system, suffer from random delays and clock deviations, which render them almost incomparable between two systems.

In so-called local broadcast networks a packet is usually received and logged by multiple systems. Using such receptions as *anchor points*, two efficient offline synchronization methods with strong theoretical properties were introduced in [3] and [2]. Both estimations are based on an affine linear clock model. As clocks built in computers are nearly affine linear under constant conditions for a short time span, this approach shows high accuracy for logs that span up to 20 minutes. Yet, for longer time spans, changing environmental conditions and non-linear clock deviations must be considered. Generalizing the mentioned approaches to a non-linear clock model is therefore a natural step.

1.1 Terminology

Let I represent a finite set of *events* and J be a finite set of *network nodes*. An event i is received by node j if and only if $(i,j) \in R \subset I \times J$. The corresponding log *timestamp* $\tau_{i,j} \in \mathbf{R}$ is based on j's local clock. Let $R^j = \{\, i \mid (i,j) \in R \,\}$ denote the events of node j and $R^{j_1,j_2} = R^{j_1} \cap R^{j_2}$ the common events of node j_1 and j_2.

It is always assumed that the timestamps are *regular* and *connected* in the following sense: for each node the timestamps of two distinct events are assumed to be distinct and the undirected graph (J,G), with

$$G = \{\, \{\, j_1, j_2 \,\} \mid j_1 \neq j_2, |R^{j_1,j_2}| \geq 2 \,\},$$

is connected. The following assumption expresses the timestamp process formally:
Main Assumption: There are increasing continuous functions $\hat{C}_j : \mathbf{R} \to \mathbf{R}$ and real numbers $\hat{s}_i$ and $\hat{d}_{i,j} \geq 0$ which generate the timestamps:

$$\hat{C}_j(\hat{s}_i + \hat{d}_{i,j}) = \tau_{i,j}.$$

We call $\hat{C}_j$ a *clock*, $\hat{s}_i$ a *synchronized timestamp* and $\hat{d}_{i,j}$ a *delay*. As a clock is increasing, for the inverse clock $\hat{c}_j = \hat{C}_j^{-1}$ this is equivalent to

$$\hat{c}_j(\tau_{i,j}) = \hat{s}_i + \hat{d}_{i,j}.$$

1.2 Basic Idea

In order to retrieve the *generating data* $\hat{c}_j, \hat{s}_i$, and $\hat{d}_{i,j}$ from $\tau_{i,j}$ only, the non-linear inverse clocks are modelled as linear combinations of some *basis functions*. It is sufficient to model an inverse clock c_j on the timestamps interval $[\tau_j, \tau_j + \Delta\tau_j]$ only, with $\tau_j = \min\{\, \tau_{i,j} \mid i \in R^j \,\}$ and $\Delta\tau_j = \max\{\, \tau_{i,j} \mid i \in R^j \,\} - \tau_j$. For

continuous basis functions $\beta_1, \ldots, \beta_n : [0,1] \to \mathbf{R}$ the inverse clocks are given as

$$c_j(\tau_j + t\Delta\tau_j) = \sum_{k=1}^{n} \beta_k(t) r_{k,j}.$$

Yet, the underdetermined equations

$$c_j(\tau_{i,j}) = s_i + d_{i,j} \qquad \forall (i,j) \in R \tag{1}$$

are not sufficient to recover the generating data. For example, shifting the synchronized timestamps (s_i) at the cost of higher delays $(d_{i,j})$ yields a new solution, without changing the inverse clocks. A "quality" measure is necessary to choose an optimal solution. Assuming the delays $(\hat{d}_{i,j})$ are independently identically exponentially distributed random variables with same mean, the *maximum likelihood estimator* can be rephrased as the following sparse linear program:

$$\min\{ \sum_{(i,j)\in R} d_{i,j} \mid d_{i,j} = c_j(\tau_{i,j}) - s_i \geq 0 \}. \tag{2}$$

Equipped with *monotonicity constraints* and *normalization* this program yields a good estimation of $\hat{c}_j, \hat{s}_i$, and $\hat{d}_{i,j}$; these estimates minimize the sum of delays. We note that the distribution assumption may not hold for real computer systems. However, experiments in [3] show that the ML estimator performs well even in this case, since it is also a constrained *least absolute deviations regressor*. It performs even better than the least squares estimator given in Section 4, although this least squares estimator is based on much weaker assumptions.

With a suitable basis the presented approach yields the same affine linear estimator as in [3]. The monotonicity constraints are theoretically important; yet, depending on the basis, they are usually not active in pratice. Some normalization is necessary, as otherwise the minimum of (2) is trivially zero and without any meaning. Finding a good normalization is not trivial; the next section will address this issue. Section 3 will give criteria on the choice of a basis and shows some examples. Consistency is discussed in Section 4. Finally some numerical results are given in Section 5.

2 Normalization

We assume that the clocks work properly within some deviation bounds, that is, they are close to the identity. Thus, estimations of the inverse clocks shall be the identity plus some offset in average. The arithmetic average $\frac{1}{|J|}\sum_{j\in J} c_j(t)$ bears no meaning since $c_{j_1}(t)$ and $c_{j_2}(t)$ refer to two different dates for distinct $j_1, j_2 \in J$. It is more intuitive to compare $c_{j_1}(\tau_{i,j_1})$ and $c_{j_2}(\tau_{i,j_2})$ for some common event $i \in R^{j_1,j_2}$. However, incomplete timestamps and random delays make this approach impracticable. Hence, the following weighted average normalization is used:

$$t = \frac{1}{|J|} \sum_{j \in J} \frac{c_j(\tau_j + t\Delta\tau_j) - c_j(\tau_j)}{\Delta\tau_j} \tag{3}$$

for all $t \in [0,1]$. Equation (3) specifies the average increase only. In addition, it is thus necessary to fix the offset $\tau_{j_0} = c_{j_0}(\tau_{j_0})$ for a $j_0 \in J$. Since c_j is a linear combination of the basis functions, the normalization reduces to linear equations in the coefficients. To satisfy the normalization the basis must be chosen accordingly.

3 Basis Functions

As stated in the preceding section, each inverse clock $\hat{c}_j$ is nearly affine linear and increasing. Thus, a reasonable basis should generate any non-decreasing affine linear function on $[0,1]$. Strict monotonicity is difficult to ensure by affine linear constraints. Hence, the monotonicity requirement is relaxed to non-decreasing. Generating affine linear functions is necessary to satisfy the normalization. To discretize the normalization the collocation matrix

$$(\beta_k(t_l))_{1 \le k,l \le n}$$

must be non-singular for some $t_1, \ldots, t_n \in [0,1]$. The basis $(\beta_1, \beta_2) = (t, 1)$, for example, yields the affine linear estimator. Requiring the coefficient of β_1 to be non-negative ensures the generated inverse clocks to be non-decreasing. Since $(t,1)$ forms a Haar system, the collocation matrix is non-singular for two distinct points.

To increase the degree of freedom, *B-splines* are a natural choice. As the timestamps give no hints on the choice of the knot sequence, uniform B-splines are a neutral and balanced option. It is well known that the collocation matrix of a B-spline system $(\beta_1, \ldots, \beta_n)$ is non-singular for $t_1, \ldots, t_n$ if the diagonal elements are positive [4]. To ensure monotonicity, it suffices if the coefficient vector is non-decreasing [1].

4 Simplified Least Squares Estimator and Consistency

Due to normalization, the estimate c_j need not be the generating inverse clock $\hat{c}_j$. Thus, s_i need not match $\hat{s}_i$. An estimation is acceptable, though, if $c_j \approx \hat{c}_j$ modulo some global transformation affecting all inverse clocks. Moreover, the chosen basis need not generate $\hat{c}_j$ at all. For these reasons, we can not expect to recover the exact data $\hat{c}_j, \hat{s}_i$ and $\hat{d}_{i,j}$. Yet suppose each exact inverse clock $\hat{c}_j$ does match the normalization and can be generated by the basis, this section shows that we can recover the exact data, even with less information.

For this section we only assume that the delays are uncorrelated with the same mean and variance σ^2 at most. Without loss of generality the following simplifying assumptions are made: the timestamps are normalized to $[0,1]$, that is $\tau_j = 0$ and $\Delta\tau_j = 1$. Let the graph (J,G) in Section 1 be a tree and for each $\{j, j'\} \in G$ let

$R^{j,j'} = R_1^{j,j'} \cup \cdots \cup R_n^{j,j'}$ be a disjoint partition of $R^{j,j'}$ with $|R_1^{j,j'}| = \cdots = |R_n^{j,j'}| = m$, that is, j and j' share mn events. Equation (1) is relaxed to

$$\frac{1}{m} \sum_{i \in R_k^{j,j'}} c_j(\tau_{i,j}) - c_{j'}(\tau_{i,j'}) = \frac{1}{m} \sum_{i \in R_k^{j,j'}} d_{i,j} - d_{i,j'}. \tag{4}$$

for $R^{j,j'} \neq \emptyset$. The average normalization is simplified to

$$c_{j_0}(t) = t \tag{5}$$

for a $j_0 \in J$ and every $t \in [0,1]$. Clearly, the generating data $\hat{c}_j$ and $\hat{d}_{i,j}$ satisfy Equation (4). For $\hat{d}_{i,j}$ the right hand side becomes a random variable with mean zero.

Let the basis functions $\beta_1, \ldots, \beta_n$ be Lipschitz continuous with constant $\frac{1}{n}L$. For $t_1, \ldots, t_n$ let the collocation matrix be non-singular and the maximum norm of the inverse be bounded by δ. Furthermore, let

$$|t_k - \tau_{i,j}| \leq \frac{\varepsilon}{\delta L}$$

for $k = 1, \ldots, n$, $i \in R_k^{j,j'}$, and some $\varepsilon \in (0,1)$. Under these assumptions Equations (4) and (5) determine the coefficients of the inverse clocks uniquely. Solving Equations (5) and (4) with zero on the right hand side in least squares sense yields a *simplified least squares estimator*.

Let r^{LS} denote the coefficient vector of all inverse clocks of the least squares estimator and $\hat{r}$ the one of all exact inverse clocks $\hat{c}_j$. Under the given assumptions

$$\|\hat{r} - r^{\mathrm{LS}}\|_\infty \leq \frac{|J|}{q^{|J|}\sqrt[3]{m}}$$

holds with probability at least $1 - 2\sigma^2/\sqrt[3]{m}$ and $q = \delta/(1-\varepsilon)$, which depends on the basis only and not on m or $|I|$. Thus, the simplified least squares estimator is consistent.

Despite this strong theoretical result, the convergence is slow in contrast to the rate of the ML estimator, which delivers more accurate results in all our numerical examples. This behaviour is discussed in [2] for the affine model.

5 Numerical Results and Conclusion

Results of experiments with synthesized and real-world data are promising. The *relative estimation error* of two nodes j, j' is approximately the delay difference:

$$d_{i,j} - d_{i,j'} = c_j(\hat{C}_j(\hat{s}_i + \hat{d}_{i,j})) - c_{j'}(\hat{C}_{j'}(\hat{s}_i + \hat{d}_{i,j'})).$$

Figure 1 shows the estimation error of a fixed clock for a given set of logs generated in a real-world experiment, which spans 100 minutes and has $6 \cdot 10^3$ events and 6 nodes. As shown in Figure 1, the cubic uniform B-spline estimator improves the estimation dramatically. The remaining error is dominated by random delays.

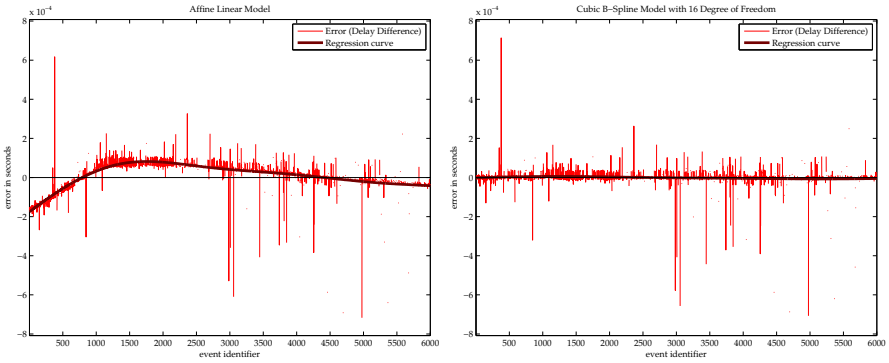


Fig. 1 Comparission between affine linear and non-linear estimator

In summary, this paper presented an efficient and theoretically sound method for offline time synchronization with a general non-linear clock model. Additional real-world experiments over longer time spans may provide a better insight into clock modelling and the choice of normalization. Since the given framework is very general, it can be easily adapted to new clock models by simply changing the basis functions.

References

1. J. M. Carnicer and J. M. Peña. Monotonicity preserving representations. In *Proceedings of the international conference on Curves and surfaces in geometric design*, pages 83–90, Natick, MA, USA, 1994. A. K. Peters, Ltd.
2. Florian Jarre, Wolfang Kiess, Martin Mauve, Magnus Roos, and B. Scheuermann. Least square timestamp synchronization for local broadcast networks. *Optimization and Engineering*, 11(1): 107–123, 2010.
3. B. Scheuermann, W. Kiess, M. Roos, F. Jarre, and M. Mauve. On the time synchronization of distributed log files in networks with local broadcast media. *IEEE/ACM Transactions on Networking*, 17(2): 431–444, 2009.
4. I. J. Schoenberg and Anne Whitney. On Polya frequency function. III. The positivity of translation determinants with an application to the interpolation problem by spline curves. *Transactions of the American Mathematical Society*, 74(2): 246–259, 1953.

Dynamic Airline Fleet Assignment and Integrated Modeling

Rainer Hoffmann

Abstract Fleet assignment is a major decision problem of the airline planning process. It assigns fleet types to scheduled flights while maximizing profit. Basic fleet assignment models (FAM) are solved irrespective of other planning steps. Although this yields an optimal assignment solution, it might cause feasibility or profitability issues in dependent planning steps. We provide an overview of strategies for integrating fleet assignment and at least one further planning step.

Traditionally, FAMs do not revise their solution if demand and thus profitability deviates from expected values. We present *dynamic fleet assignment* which comprises both assigning fleet types based on expected values of demand and demand-driven re-fleeting. Further, we incorporate robustness into the FAM which increases re-fleeting flexibility. Additionally, we propose an extension of an existing robustness concept. A dynamic fleet assignment process for a medium-sized schedule is simulated in computational experiments. The results demonstrate that for a large demand variation using a robust fleet assignment leads to higher re-fleeting profits than applying a non-robust model. Further, our proposed robustness strategy largely outperforms the traditional robustness concept.

1 Introduction

The fleet assignment problem seeks to assign fleet types to flight legs such that capacity matches demand as close as possible for every flight. That is associated with the following trade-off: If an aircraft is too small to accommodate all customers, a fraction of passengers is spilled which implies lost revenue. On the other hand, an aircraft whose capacity exceeds demand features higher operating cost than a smaller aircraft with sufficient capacity.

Rainer Hoffmann
Karlsruhe Institute of Technology, Institute of Operations Research, Schlossbezirk 14, 76131 Karlsruhe, e-mail: ra.hoffmann@kit.edu

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 635
DOI 10.1007/978-3-642-20009-0_100,

Usually, the fleet assignment decision is modeled by using a multi-commodity time-expanded network flow problem. The network, as depicted in Figure 1, is for-

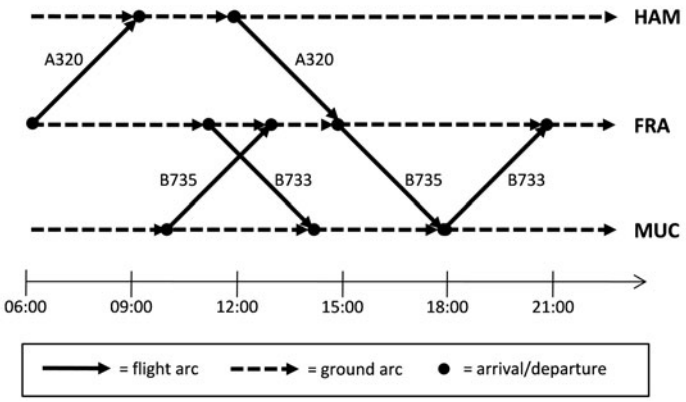


Fig. 1 Time-Space Network with possible assignment

mulated as a mixed integer problem, and a solution is obtained by applying standard solution techniques like branch-and-bound. The problem formulation as proposed by [3] is referred to as *basic* FAM and features an objective function that maximizes operational profits subject to the following constraints: First, each flight leg in the schedule has to be covered once (*cover constraint*). Second, it is necessary to ensure conservation of aircraft flow, that is, no aircraft can leave without having arrived and vice versa (*balance constraint*). Third, the number of available aircraft per fleet type must not be exceeded (*count constraint*).

2 Integrated Modeling

The basic FAM is traditionally solved irrespective of other planning steps like maintenance routing or crew scheduling. For this reason, an assignment might be optimal, but it might also cause feasibility issues in subsequent planning steps. For example, an assignment that does not provide sufficient time windows for maintenance makes a subsequent manual modification of the assignment necessary. Further, optimizing planning problems sequentially might lead to a lower overall profit than a simultaneous approach since the solutions of the planning steps are not independent from each other. In order to achieve a robust and profit-maximizing solution, other planning steps have to be integrated into the fleet assignment problem so that a solution can be determined simultaneously. However, integrated models are more complex due to an increased number of decision variables and side constraints which implies significant computational issues. Thus, new solution strategies are required to solve integrated problems in a reasonable amount of time (cf. [5] or [4]).

FAMs that integrate schedule design (e.g. [5]) either allow for varying departure times within time windows, or they develop a schedule and perform a fleet assignment simultaneously. One advantage of such an integrated model is that flight connections can be made possible by adjusting departure times. This results in a lower number of required aircraft.

Models that consider maintenance routing in the fleet assignment problem (e.g. [1]) assign fleet types to predefined flight routes while satisfying maintenance requirements for each individual aircraft. In a sequential approach it might happen that the fleet assignment is inconsistent with scheduled maintenance opportunities. For instance, a fleet type is scheduled to be maintained at a particular station but overnights at a different station that cannot provide maintenance. Therefore, a simultaneous solution of fleet assignment and maintenance routing is necessary to support compliance with safety regulations without modifying the assignment manually.

FAMs accounting for crew scheduling (e.g. [2]) usually assign fleet types to flight legs and allocate crews to a sequence of legs. Since a crew's time-away-from-base is a major determinant of crew cost, models seek to reduce waiting time between flights during the day and waiting time after overnight rests. Integrated models adjust the assignment such that the overall profit, which considers crew cost, is maximized.

3 Dynamic Fleet Assignment

FAMs generally determine a solution once and never revise it again. However, a situation might occur where a change in fleet assignment is reasonable. Since the assignment decision is based on demand forecasts that are performed several months prior to departure, deviations from expected demand values will most likely occur. In this case a modification of the assignment according to updated booking data leads to increased revenues. A further benefit from adapting a fleet assignment to actual demand is that revenue management performs best if available capacity is as close as possible to what is truly required (cf. [6]). *Dynamic fleet assignment* captures this issue and optimizes the fleet assignment by going through the following two sequential steps (as depicted in Figure 2):

1. several months prior to departure a fleet assignment decision is made based on expected values of demand;
2. demand-driven modifications of the assignment are performed close to operations (close-in re-fleeting).

We consider irregular deviations from demand forecasts which means that a modification may be performed only once, and the original assignment needs to be reconstituted the next day. In order to account for one-time modifications in fleet assignment, we propose a daily demand-driven re-fleeting model which is capable of preserving the same distribution of fleet types across the network at the beginning

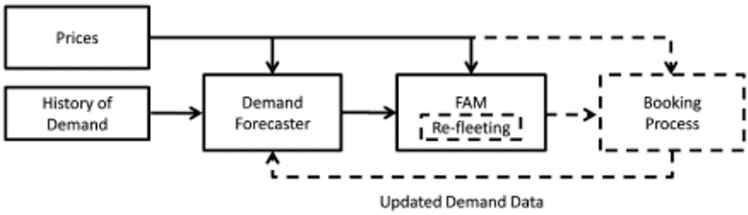


Fig. 2 Dynamic fleet assignment; dashed lines represent additional steps in the second stage

of each day.

Crew schedules are mostly determined several weeks prior to departure and cannot be changed any more thereafter. This rules out swapping aircraft that cannot be operated by the same crew. However, since crews are allowed to fly every fleet type of a crew-compatible fleet family, re-fleeting within families is still possible. The number of swapping opportunities within families is referred to as re-fleeting flexibility and determines the success of re-fleeting significantly.

Increasing re-fleeting flexibility emerged in the field of schedule recovery. In this context *robustness* is applied in order to facilitate re-fleeting in case of operational disruptions. Robustness decreases the dispersion of fleet types and families across the network which yields a concentration of flights of the same fleet type/family at particular stations; this provides an increased number of swapping opportunities. Moreover, a reduction of combinations of fleet types and stations implies cost savings through economies of scope. Therefore, robustness leads to both an increased number of swapping opportunities and decreased operating cost.

According to [6] demand variation can be considered as an operational disruption as well. Accordingly, robustness can be applied in dynamic fleet assignment like in schedule recovery in order to increase re-fleeting flexibility. To proof this supposition, we experimentally test the following hypothesis:

Applying a robust FAM in the first step of dynamic fleet assignment increases the number of swapping opportunities, and thus a higher close-in re-fleeting profit will be achieved in case of variations in demand.

We apply *station purity* in our experiments as one way of integrating robustness into the FAM. This approach, which was proposed by [7], limits the number of fleet types/families serving a particular station (*purity level*). Moreover, we extend this concept and propose *network purity* which imposes a limit on the total number of combinations of fleet types and stations in the entire network. The fundamental idea of this model is that each station's optimal purity level should be determined during the optimization process rather than imposing it ex ante. Therefore, stations do not feature a predefined purity level but rather an *average* purity level resulting from the assignment of fleet/station combinations.

4 Computational Experiments

Our data set comprises a generic schedule containing 700 flights in a hub-and-spoke network (retrieved from Deutsche Lufthansa's winter 2009/2010 schedule), 86 stations (2 hubs, 84 spokes), 16 fleet types, and 8 fleet families. Based on this data set we simulate a dynamic fleet assignment process in two subsequent stages:

1. a fleet assignment is performed based on *expected values of demand*; for each of the following models an assignment is determined: (i) a basic FAM as proposed by [3], (ii) a robust FAM incorporating station purity (*SPFAM*), and (iii) a robust FAM incorporating network purity (*NPFAM*);
2. *actual demand* on each leg is simulated, and the schedule is re-fleeted within each fleet family (based on the assignments from stage 1); a sufficiently large number of simulation runs ensures statistical significance of the results.

The results of the fleet assignments based on expected values show that the robust model featuring the best result (NPFAM) performed almost equally as FAM in terms of expected profit and load factor (difference of less than 0.2%). On the other hand, when demand deviates from expected values, as simulated in the second step, re-fleeting results point out that for a small demand variation incorporating robustness into the initial fleet assignment decreased the profit gap to FAM; for a large demand variation (coefficient of variation $\geq$ 50%) robust FAMs even demonstrated a superiority over FAM in terms of profit and load factor. Thus, we experimentally proofed our hypothesis to be true for large demand variations. The profit performance of robust models relative to FAM is depicted in Figure 3. The figure

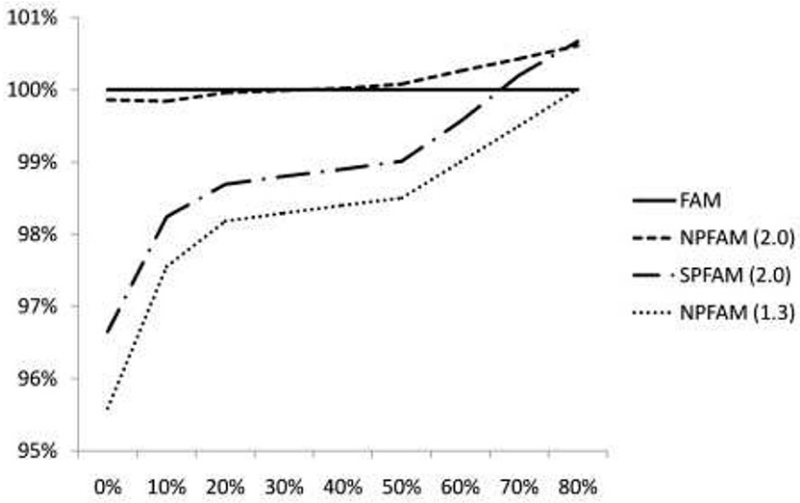


Fig. 3 Profit performance of NPFAM and SPFAM relative to FAM for different coefficients of variation (horizontal axis). The number in parentheses denotes the imposed (average) purity level.

shows that NPFAM featuring a purity level of 2 achieved the highest profit for a large coefficient of variation. The figure also shows that increasing station purity did not lead to higher profits. The robust model featuring an average purity level of 1.3 performed significantly worse than FAM for small and medium coefficients of variation. Thus, a profit-maximizing purity level has to be determined, which can be achieved, for instance, through simulation.

Our proposed approach of incorporating robustness into the FAM by imposing network purity performed generally better than SPFAM. The following observations were made for the assignments based on expected demand values: First, NPFAM required more than 80% ($\widehat{=}$ 15 hours) less run time than SPFAM to solve the FAM. Second, NPFAM achieved a 2.5% larger expected profit and a 0.5 percentage points larger expected load factor than SPFAM. Third, both models featured an equal number of combinations of fleet types and stations, but NPFAM's solution comprised significantly fewer combinations with only one arrival and one departure of a fleet family. The re-fleeting results based on varying demand show that NPFAM was superior to SPFAM in terms of profit, load factor, and spill cost for small and large demand variations; SPFAM achieved better results only for an extreme demand variation (coefficient of variation $\geq$ 80%), as depicted in Figure 3.

References

1. Cynthia Barnhart, Natashia L. Boland, Lloyd W. Clarke, Ellis L. Johnson, George L. Nemhauser, and Rajesh G. Shenoi. Flight string models for aircraft fleeting and routing. *Transportation Science*, 32(3): 208–220, 1998.
2. Chunhua Gao, Ellis Johnson, and Barry Smith. Integrated airline fleet and crew robust planning. *Transportation Science*, 43(1): 2–16, 2009.
3. Christopher A. Hane, Cynthia Barnhart, Ellis L. Johnson, Roy E. Marsten, George L. Nemhauser, and Gabriele Sigismondi. The fleet assignment problem: solving a large-scale integer program. *Mathematical Programming*, 70(2): 211–232, 1995.
4. Manoj Lohatepanont and Cynthia Barnhart. Airline schedule planning: Integrated models and algorithms for schedule design and fleet assignment. *Transportation Science*, 38(1): 19–32, 2004.
5. Brian Rexing, Cynthia Barnhart, and Tim Kniker. Airline fleet assignment with time windows. *Transportation Science*, 34(1): 1–20, 2000.
6. Sergey Shebalov. Practical overview of demand-driven dispatch. *Journal of Revenue and Pricing Management*, 8(2–3): 166–173, 2009.
7. Barry C. Smith and Ellis L. Johnson. Robust airline fleet assignment: Imposing station purity using station decomposition. *Transportation Science*, 40(4): 497–516, 2006.

Solving Multi-Level Capacitated Lot Sizing Problems via a Fix-and-Optimize Approach

Florian Sahling

Abstract We present a solution approach for the dynamic multi-level capacitated lot sizing problem (MLCLSP) and its extensions. The objective is to determine a cost minimizing production plan for discrete products on multiple resources. The time-varying demand is assumed to be given for each product in each period and has to be completely fulfilled. The production is located on capacity constrained resources for the different production stages. In an iterative fashion, our Fix-and-Optimize approach solves a series of mixed-integer programs. In each of these programs all real-valued variables are treated, but only a small and iteration-specific subset of binary setup variables is optimized. All remaining binary variables are fixed. A numerical study shows that the algorithm provides high-quality results and that the computational effort is moderate.

1 Introduction

A specific setup is often required to prepare a machine for the production of a specific product, if this machine produces different product types. Whenever this changeover causes setup times and/or cost, a lot sizing problem arises. In this paper, we consider the multi-level capacitated lot sizing problem (MLCLSP), see [1]. A flexible solution approach is presented for the MLCLSP based on mathematical programming. In this so-called Fix-and-Optimize heuristic, a series of mixed-integer programs (MIPs) derived from the MLCLSP is solved in an iterative way. In each MIP all real-valued decision variables and only a subset of the binary setup variables are solved to optimality. These binary setup variables are fixed afterwards to their optimal values. The following MIP includes another subset of the binary variables which are solved to optimality.

Institut für Produktionswirtschaft
Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, e-mail: florian.sahling@prod.uni-hannover.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 641
DOI 10.1007/978-3-642-20009-0_101,

2 Problem Statement and Model Formulation

The objective of the MLCLSP is to determine a production plan including production quantities Q_{kt} and end-of-period inventory levels Y_{kt} for all products k over the complete planning horizon $(t = 1, \ldots, T)$. This production plan should minimize the sum of setup, holding and overtime cost. The primary demand d_{kt} is known for each product k and period t in advance and has to be fulfilled in time. The production of a product during a period requires a setup which leads to both setup time and setup cost. The setup state $(\gamma_{kt} \in \{0,1\})$ of a specific product is lost at the end of a period. The given capacity b_{jt} of resource j in period t can be extended by the use of overtime O_{jt}. The MLCLSP can be stated formally with the notation in Table 1.

Table 1 Notation for the MLCLSP

Sets:	
$k, i \in \mathcal{K}$	products
$t \in \mathcal{T}$	periods
$j \in \mathcal{J}$	resources
$\mathcal{K}_j$	set of products requiring resource j
$\mathcal{N}_k$	set of immediate successors of product k
Parameters:	
a_{ki}	number of units of product k required to produce one unit of product i
b_{jt}	available capacity of resource j in period t
B	big number
d_{kt}	external demand of product k in period t
h_k	holding cost of product k per unit and period
oc_{jt}	overtime cost per unit of overtime at resource j in period t
s_k	setup cost of product k
tp_k	production time per unit of product k
ts_k	setup time of product k
z_k	planned lead time of product k
Decision variables:	
O_{jt}	overtime at resource j in period t
Q_{kt}	production quantity (lot size) of product k in period t
Y_{kt}	planned end-of-period inventory of product k in period t
γ_{kt}	binary setup variable of product k in period t

The MLCLSP can then be formulated as a MIP, see [1].

Model MLCLSP

$$\text{Minimize } Z = \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} (s_k \cdot \gamma_{kt} + h_k \cdot Y_{kt}) + \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} oc_{jt} \cdot O_{jt} \tag{1}$$

subject to:

$$Y_{k,t-1} + Q_{k,t-z_k} - \sum_{i \in \mathcal{N}_k} a_{ki} \cdot Q_{it} - Y_{kt} = d_{kt} \qquad \forall k,t \qquad (2)$$

$$\sum_{k \in \mathcal{K}_j} (tp_k \cdot Q_{kt} + ts_k \cdot \gamma_{kt}) \leq b_{jt} + O_{jt} \qquad \forall j,t \qquad (3)$$

$$Q_{kt} - B \cdot \gamma_{kt} \leq 0 \qquad \forall k,t \qquad (4)$$

$$Q_{kt}, Y_{kt} \geq 0 \qquad \forall k,t \qquad (5)$$

$$O_{jt} \geq 0 \qquad \forall j,t \qquad (6)$$

$$\gamma_{kt} \in \{0,1\} \qquad \forall k,t \qquad (7)$$

The objective function (1) minimizes the sum of setup, holding and overtime cost. The inventory balance equations (2) ensure that the primary and secondary demand is met. Note that the lead time z_k is necessary to ensure that is possible to disaggregate this production plan into a feasible machine schedule, see [4].

Inequalities (3) state that production quantities and setups must meet the capacity constraints for all the resources. The inequalities (4) link the decision variables for the production quantities with the binary setup variables. In the case of a production $(Q_{kt} > 0)$, the respective setup variable γ_{kt} is forced to be 1. The decision variables for the production quantities and end-of-period inventory levels (5) as well as for the planned overtime (6) cannot be negative.

3 An Iterative Optimization-Based Heuristic

3.1 Basic Idea of the Fix-and-Optimize Heuristic

[3] have shown that the MLCLSP is $\mathcal{NP}$-hard. Furthermore, the question whether a feasible production plan (without overtime) exists has been shown to be $\mathcal{NP}$-complete in the case of positive setup times, see [5]. Thus, the solution times are prohibitively large even for small problem instances. Several authors have therefore developed heuristics for the MLCLSP. An extensive review of the literature on solution approaches for capacitated lot sizing problems is given amongst others by [2]. Most of the computational effort is required for solving the binary setup variables optimally, while the solution time is almost negligible for the real-valued variables for a given setup pattern. The basic idea of our Fix-and-Optimize heuristic (see [4] and [6]) is to solve a sequence of subproblems derived from the MLCLSP in an iterative fashion. In each subproblem, the number of "free" binary setup variables γ_{kt} is limited as most of the binary setup variables γ_{kt} are fixed to a constant setup state $\overline{\gamma}_{kt}^{fix}$. The resulting subproblem can be solved to optimality quickly. This (optimal) solution of the subproblem describes a new temporary solution for the binary setup variables. At least some of them are fixed in the next subproblem when a new sub-

set of binary variables is solved to optimality. In contrast, the real-valued decision variables are never fixed for all products, periods and machines.

3.2 Definition of Subsets of Binary Setup Variables

The numerical effort to solve each subproblem exactly depends on the number of binary variables considered simultaneously. An increasing number of "free" binary variables solved to optimality in each subproblem raises the required solution time. However, at the same time the solution quality of the Fix-and-Optimize heuristic will increase as well. Therefore, the definition of subproblems is important. We defined three different strategies to identify the binary variables considered in the current subproblem. Each strategy takes trade-offs between the products into account (see [4, pp. 250–251]).

- **Product-oriented decomposition:** Each subproblem corresponds to a single product k. In each subproblem, all setup variables are optimized for the respective product over the complete planning horizon.
- **Resource-oriented decomposition:** Each subproblem corresponds to a single resource j and a subset of periods t. The subset of periods contains λ successive periods. Two successive subproblems related to the same resource j have an overlap of θ periods. In each subproblem, all products $k \in \mathcal{K}_j$ requiring resource j are considered.
- **Process-oriented decomposition:** Each subproblem corresponds to a subset of periods and a direct predecessor-successor-relationship between a product k and one of its immediate successors $i \in \mathcal{N}_k$. For each direct predecessor-successor-relationship two subproblems are defined. The first one covers the first half of the planning horizon and the other one the second half.

3.3 Iterative Algorithm

Initially, we start with a trivial solution $(\overline{\gamma}_{kt}^{fix} = 1, \forall k,t)$. This solution is improved iteratively while unattractive setup decisions are eliminated very quickly. The decomposition strategies are combined to variants. In each variant the decomposition strategies are executed consecutively, but each decomposition strategy is executed once in each iteration of the Fix-and-Optimize heuristic. Each variant can either be treated just once or repeated until a local optimum is reached. In each variant we start with a product-oriented decomposition.

4 Numerical Results

In extensive numerical studies based on test sets proposed by [9], [6] has shown that the Fix-and-Optimize heuristic provides high quality results for the MLCLSP with respect to both solution time and solution quality. We have compared our results to two well-known solution approaches in literature, namely the Lagrangean relaxation approach by [11] (TDH) and the time-oriented decomposition approach by [8] (SH). Compared to TDH our Fix-and-Optimize heuristic is superior in terms of solution quality, while the TDH is so fast that the solution time can be neglected. Although the SH offers better results than the TDH, the Fix-and-Optimize heuristic outperforms the SH with respect to both solution time and quality.

5 Outlook

The Fix-and-Optimize heuristic has been successfully adapted to model extensions and modifications. [7] extended the Fix-and-Optimize heuristic to solve dynamic multi-level lot sizing problems with linked lot sizes (MLCLSP-L, see [10]). In the MLCLSP-L, the setup state of a resource can be carried over to subsequent periods. To solve the MLCLSP-L, only the definition of subproblems has to be adjusted. The solution quality is comparable to the one obtained from the MLCLSP. The MLCLSP with sequence dependent setup times and cost can be solved as well by the Fix-and-Optimize heuristic, see [6]. This shows the high degree on flexibility of the Fix-and-Optimize heuristic. Due to this flexibility the Fix-and-Optimize heuristic seems to be a promising candidate to solve other combinatorial optimization problems.

References

1. Billington PJ, McClain JO, Thomas LJ (1983) Mathematical programming approaches to capacity-constrained MRP systems: Review, formulation and problem reduction. Management Science 39: 1126–1141
2. Buschkühl L, Sahling F, Helber S, Tempelmeier H (2010) Dynamic capacitated lot-sizing problems – a classification and review of solution approaches. Operations Research Spectrum 32: 231–261
3. Florian M, Lenstra JK, Rinnooy Kan AHG (1980) Deterministic production planning: Algorithms and complexity. Management Science 26: 669–679
4. Helber S, Sahling F (2010) A fix-and-optimize approach for the multi-level capacitated lot sizing problem. International Journal of Production Economics 123: 247–256
5. Maes J, McClain JO, van Wassenhove LN (1991) Multilevel capacitated lotsizing complexity and LP-based heuristics. European Journal of Operational Research 53: 131–148
6. Sahling F (2010) Mehrstufige Losgrößenplanung bei Kapazitätsrestriktionen. Gabler, Wiesbaden
7. Sahling F, Buschkühl L, Helber S, Tempelmeier H (2009) Solving a multi-level capacitated lot sizing problem with multi-period setup carry-over via a fix-and-optimize heuristic. Computers & Operations Research 36: 2546–2553

8. Stadtler H (2003) Multilevel lot sizing with setup times and multiple constrained resources: Internally rolling schedules with lot-sizing windows. Operations Research 51: 487–502
9. Stadtler H, Sürie C (2000) Description of MLCLSP test instances. Tech. rep., Technische Universität Darmstadt
10. Sürie C, Stadtler H (2003) The capacitated lot-sizing problem with linked lot sizes. Management Science 49: 1039–1054
11. Tempelmeier H, Derstroff M (1996) A lagrangean-based heuristic for dynamic multilevel multiitem constrained lotsizing with setup times. Management Science 42: 738–757

Online Optimization: Probabilistic Analysis and Algorithm Engineering*

Benjamin Hiller

Abstract This article gives an overview on some of the results of the authors' PhD thesis [3]. The subject of this thesis is *online optimization*, which deals with making decisions in an environment where the data describing the process to optimize becomes available over time, i. e., *online*. In particular, we study algorithms for *combinatorial* online optimization problems involving discrete decisions both from a practical and a theoretical point of view. Here we sketch our results related to the control of elevators in high-rise buildings.

1 Introduction

Online optimization problems occur frequently in practice. The most prominent example is probably the control of elevators: Passengers arrive over time, registering their travel calls. The elevator control now needs to update the schedule for serving all currently known travel calls *online*, i. e., immediately, without waiting for future travel calls. In addition, the new schedule needs to be computed in less than a second, i. e., in *real time*. The overall goal is to achieve small waiting times for all passengers.

In the first part of the thesis, summarized in Section 2, we develop new elevator control algorithms for *destination call systems*, in which passengers register their destination floor already at the start floor. The algorithms we propose are *reoptimization algorithms* which, each time the schedule needs to be updated, try to compute a schedule with small waiting times for the currently known travel calls. We focus on *exact* reoptimization algorithms, i. e., ones that are able to find provably optimal schedules. Since even schedules which are optimal for the current situation

Zuse Institute Berlin, Takustraße 7, D–14195 Berlin,Germany, hiller@zib.de

* Supported by the DFG research group "Algorithms, Structure, Randomness" (Grant number GR 883/10-3, GR 883/10-4) and a DAAD dissertation grant.

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_102,

may turn out inappropriate in view of the future system evolution it is not clear that a reoptimization algorithm provides a good online control. We do not incorporate models of future travel calls in our algorithm, since on the one hand the computational complexity would increase significantly and on the other hand it is not clear how to model them appropriately, i. e., how they are distributed over space and time. However, we also study means to change the reoptimization problem such that certain effects on future travel calls are taken into account. Our algorithms have in part been implemented by Kollmorgen Steuerungstechnik, our industry partner.

The second, theoretical part of the thesis addresses the question on how to judge theoretically whether an online algorithm performs well. Ideally, we would like to have "optimal" online algorithms, but a mathematically very interesting aspect of online optimization is that there is no unique or canonical definition of what "optimal online algorithm" actually means. Roughly speaking, the issue is that there are many possible "futures" and in general it is not possible to have an algorithm that is best for all "futures". One therefore needs to define how to compare algorithm results on different "futures". From a pragmatic and practical point of view one compares online algorithms by simulating them on relatively few futures, considering an algorithm to be optimal if one has not found a better one. Many real-world problems (elevator control as well) are so complex that we cannot provide better answers yet. For simpler problems it is however possible to prove results for certain definitions of "optimality". A new definition of "optimality" as well as first results are outlined in Section 3.

2 Elevator Control for Destination Call System

The elevator control problem may be viewed as a special vehicle routing problem. However, elevator schedules are much more complex than the vehicle tours considered there. The main reasons for this are the high capacity allowing many passengers to be served simultaneously and the involved sequencing constraints for the passengers carried. For instance, it is required that a passenger needs to be transported to his destination floor before the elevator may change its direction.

Cooperating with Kollmorgen Steuerungstechnik, we first developed a heuristic, i. e., non-exact, reoptimization algorithm BestInsert, which is real-time compliant on microcontrollers. This algorithm has been implemented by Kollmorgen Steuerungstechnik and is now running in several installations world-wide. To assess whether the algorithm may be improved by a stronger reoptimization we developed an exact reoptimization algorithm ExactReplan. This algorithm is based on a set partitioning model that is solved using column generation, generating columns using a Branch & Bound algorithm. For the Branch & Bound algorithm, we developed new lower bounds. As far as we know, ExactReplan is the first exact reoptimization algorithm controlling a group of passenger elevators.

All destination call systems implemented so far respond to a registered destination call by assigning the passenger to a serving elevator. The passenger may then

proceed to this elevator immediately. Such *immediate assignment (IA)* systems have the disadvantage that the assignment of passengers to elevators is fixed very early and cannot be changed later on. We therefore also studied (hypothetical) *delayed assignment (DA)* systems, in which an elevator signals the served destination floors shortly before its arrival, thus selecting the passengers to serve. This allows to defer the assignment decision as long as possible. From a passenger point of view a DA system works analogously to a conventional 2-button system, the difference being that passengers and elevator control now communicate destination floors instead of travel directions.

Using ExactReplan in extensive simulations it was possible to compare the relative performance of IA and DA systems. The performance-critical traffic situation in office buildings is *(morning) up peak traffic* [2], during which many passengers need to be transported to their office floors in a short time span. Particularly important is the so-called *handling capacity (HC)*, measuring the ratio of the building population that can be served with acceptable waiting times in five minutes. A HC of 14 % is considered very good. Although the DA system achieved shorter waiting times than the IA system throughout our simulations, it does not achieve a higher HC if the reoptimization is based on service quality alone. We also studied a restriction of the reoptimization model that enforces a stronger grouping of passengers by destination floors. This restriction allows to improve the HC further. Moreover, the DA system now achieves a 1 % higher HC than an IA system, i. e., the HC improves from 14 % to 15 % and from 16 % to 17 % in the two buildings studied.

The following results are the most relevant for practice. First of all, ExactReplan allows an evolution of our heuristic algorithm BestInsert that is in use today. We found that BestInsert performs in an IA system almost as good as ExactReplan, whereas ExactReplan achieves shorter waiting times in a DA system. Moreover, the HC of a system controlled by BestInsert is at least 50 % higher than if a conventional 2-button control is applied. Thus it is possible to transport at least one and a half as many passengers with acceptable waiting time by changing from a conventional control to a destination call control.

3 A New Approach for Analyzing Online Algorithms

The most common definition of "optimality" for online algorithms is based on the *competitive ratio*. To this end one considers a hypothetical clairvoyant algorithm OPT, which in contrast to the online algorithm knows the entire future and is able to determine a cost-optimal solution for it. Every possible future yields a ratio of the cost of the online algorithm to the cost of OPT. The competitive ratio is now the maximum of this ratio for all possible futures. It measures how much worse than OPT an online algorithm may perform. Two online algorithms ALG_1 and ALG_2 can now be compared via their competitive ratios. An algorithm is optimal if it achieves the least possible competitive ratio.

	ALG_1	ALG_2	OPT
σ_1	4 1.33	5 1.6	3
σ_2	3 1.5	3 1.5	2
σ_3	3 1	4 1.33	3
σ_4	4 2	3 1.5	2
σ_5	2 1	2 1	2

Fig. 1 Comparison of algorithms using their competitive ratio.

Regarding the comparison of all possible futures comparing the competitive ratios means that the worst future for ALG_1 is compared to the worst future of ALG_2. An example is shown in Figure 1, indicating the cost of ALG_1, ALG_2, and OPT, respectively, on five possible futures $\sigma_1, \ldots, \sigma_5$. In addition the ratio of algorithm cost to the cost of OPT is shown; it can be seen that the competitive ratio of ALG_1 is determined by future σ_4, whereas that of ALG_2 is determined by future σ_1. Obviously, ALG_1 is clearly inferior to ALG_2 in terms of the competitive ratio. However, looking at the cost more closely reveals that ALG_1 is better than ALG_2 on futures σ_1, σ_3, worse on σ_4 and as good as ALG_2 on σ_2, σ_5. We can see that it is possible that ALG_1 is worse than ALG_2 w. r. t. the competitive ratio, despite yielding the same or better cost on almost every future.

This property is due to the worst-case nature of the competitive ratio and has surprising consequences for some online optimization problems. This shows up prominently for the online bin coloring problem [4]. The task is to pack colored items into bins, such that the number of distinct colors in a bin is as small as possible. At most m bins may be used at the same time; each bin provides capacity for exactly B items. As an item and its color become known, the item needs to be packed irrevocably into one bin. Only then the next item becomes known. If a bin is filled to capacity, it is replaced by an empty one. A natural algorithm for this problem is GreedyFit (see Figure 2): If the color of the new item is present in one of the bins, GreedyFit puts the item into this bin. Otherwise a bin with the least number of distinct colors is selected. This algorithm is *greedy*, since it keeps the number of distinct colors as small as possible in each step. An alternative algorithm is OneBin: Put all items in the first bin, disregarding their color. Krumke et al. [4] showed that OneBin has a strictly better competitive ratio than GreedyFit, although it uses the available resources less efficiently. However, simulation on random color sequences indicates that GreedyFit is almost always better than OneBin.

We developed a new analytical approach allowing us to show that GreedyFit is "almost always" better than OneBin, which gives a more realistic picture of the relative performance than the competitive ratio. We assume that the sequence of item colors is generated randomly. Thus the maximal number of distinct colors an algorithm puts in a bin becomes a random variable. Our approach compares two algorithms by comparing these random variables using stochastic orders, which generalize partial orders. In particular, we use the stochastic dominance order, showing that the number of distinct colors of GreedyFit is stochastically dominated by that of OneBin, i. e., OneBin puts in a stochastic sense more colors in a bin.

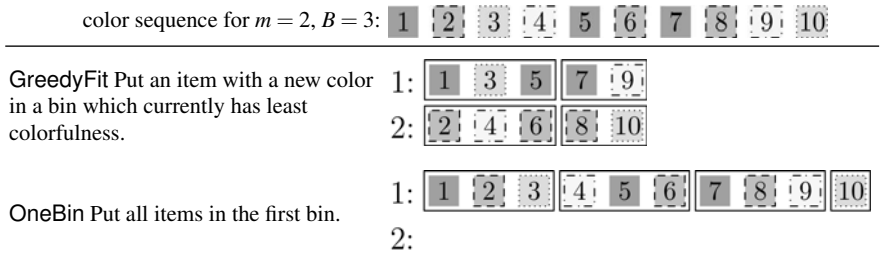


Fig. 2 Algorithms GreedyFit and OneBin on an example color sequence. For improved recognizability each color has an additional dash pattern.

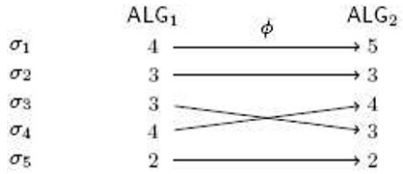


Fig. 3 Illustration of the idea of bijective analysis for the example from Figure 1: ALG_1 is at least as good as ALG_2 along the bijective mapping ϕ.

An important special case of our stochastic dominance analysis is bijective analysis [1], which was developed simultaneously. Bijective analysis compares online algorithms by "charging" possible futures with others in the following way. One says that ALG_1 is better than ALG_2, if possible futures can be mapped bijectively onto themselves such that ALG_1 performs on the original future as least as good as ALG_2 on the image future (see Figure 3). As for the competitive ratio, different futures are compared with each other; however, instead of comparing worst-case futures only the required bijectivity takes *all* possible futures into account. The hypothetical and unrealistic algorithm OPT is not needed any more. Our stochastic dominance result implies a bijective analysis result, but is much stronger.

We conclude by describing how the online bin coloring problem is connected to elevator control and what can be learned from it for this application. As already mentioned, up peak traffic requires to transport many passengers from the entrance floor to their office floors in a short time period. If the elevator control algorithm generates a schedule where the elevators need a long time to return to the entrance floor, future passengers arriving there need to wait long. For this reason the *roundtrip time*, i. e., the time between two successive arrivals of one elevator at the entrance floor, should not be too long. It is clear that the roundtrip time increases with the number of stopping floors. In a conventional system, in which a passenger registers first his travel direction on the start floor and then, after the elevator has arrived, his destination floor, the elevator control has no means to influence which passengers use which elevator. In particular, it cannot group passengers by their destination floor to save stops. In contrast, a destination call system can do this grouping to achieve short roundtrip times.

This is the online bin coloring problem in disguise. Considering passengers as items, their destination floors as colors, and the elevators as bins, OneBin corresponds to a conventional system and GreedyFit to a variant of a destination call system: As a conventional system cannot direct passengers, all passengers try to use the next elevator (first bin), which then travels to many floors (distinct colors). A destination call system on the other hand may group passengers by destination floors (colors) and assign them to elevators (bins) such that each of them has few stopping floors (contains distinct colors). Applied to this context, our stochastic dominance result states that a destination call system achieves short roundtrip times for a larger share of possible futures (future transportation calls) than a conventional system. Grouping by destination floors occurs to some extent if reoptimization is based on service quality only. For higher traffic intensities the grouping may be enforced more strongly by restricting the reoptimization model as already mentioned. This restriction was in fact motivated by our online bin coloring analysis.

References

1. S. Angelopoulos, R. Dorrigiv, and A. López-Ortiz. On the separation and equivalence of paging strategies. In *SODA 2007*, pages 229–237, 2007.
2. G. C. Barney. *Elevator Traffic Handbook: Theory and Practice*. Taylor and Francis, 2002.
3. B. Hiller. *Online Optimization: Probabilistic Analysis and Algorithm Engineering*. PhD thesis, TU Berlin, 2009.
4. S. O. Krumke, W. E. de Paepe, L. Stougie, and J. Rambau. Bincoloring. *Theoret. Comput. Sci.*, 407(1–3): 231–241, 2008.

Solving Stochastic Energy Production Problems by a Scenario Tree-Based Decomposition Approach

Debora Mahlke

Abstract This paper is concerned with the development of an optimization method for multistage stochastic mixed-integer programs arising in energy production. This novel method relies on the decomposition of the original problem formulation into several subproblems based on splitting the scenario tree into subtrees. On this basis, a branch-and-bound method is applied in order to recover feasibility and thus to solve the problem to global optimality. As an application, we present a power generation problem with fluctuating wind power supply in order to investigate the potential of energy storages to decouple fluctuating supply and demand.

1 Introduction

Power generation based on renewable energy sources plays an important role in the development of a sustainable generation of electrical energy. In particular, wind energy is considered to be most promising to provide a substantial part of the electrical energy supply. But due to the fluctuating behavior of power production from renewable energies, especially caused by wind power production, new challenges are posed to the structure of power generation systems.

The increasing feed-in of fluctuating power into the electricity grid influences the operating requirements of the conventional power plants leading to a rising participation of these plants to the balance energy. The additional necessity of regulation affects the efficiency of power plants, as the generation efficiency strongly depends on the current production level.

Apart from the fluctuating power supply, further stochastic aspects arise from the uncertain electricity prices. These prices have a major impact on the decision about the commitment of energy storages and are hardly predictable since the liberalization of the German energy market.

Debora Mahlke
Technische Universität Darmstadt, e-mail: mahlke@mathematik.tu-darmstadt.de

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 653
DOI 10.1007/978-3-642-20009-0_103, © Springer-Verlag Berlin Heidelberg, 2011

A possibility of further increases in wind energy is the commitment of energy storages in the generation system. Although to a limited extent, energy storages are capable to convert and store surplus energy when produced and to supply energy in times of peak demand. By transforming base load capacity to peak load capacity, they thus can contribute to prevent partial load operation of conventional power plants. Under consideration of these facts, we face the question of how energy storages can contribute to decouple demand and supply if large amounts of fluctuating regenerative energy are integrated into a power generation network.

In order to reasonably account for the uncertainty arising within the problem described above, specific models and solution methods are necessary. In case probabilistic information of the unknown parameters is available, stochastic programming provides a useful framework in order to integrate the uncertainty into the model. However, from the computational point of view the solution of the resulting multistage stochastic programs still poses a great challenge. If additionally integer restrictions are integrated, as in the problem at hand, novel solution approaches become necessary in order to solve problems of practical relevant sizes. This motivates our studies which focus on the development of suitable solution methods for multistage stochastic mixed-integer programs with the aim of of solving the application problem specified subsequently.

2 An Energy Production Problem

In order to analyze the potential of energy storages within a power generation system including fluctuation energy supply, we formulate a multistage stochastic mixed-integer optimization problem. To receive reliable results, we aim at a formulation sufficiently close to reality. In particular, the model states the basic connections within the power generation system, but also takes technical and economical aspects of the production units and storages into account. For the optimization, we are interested in the consideration of a predefined planning horizon comprising several days. This span of time is discretized into subintervals of 15 minutes permitting the consideration of the relevant technical behavior of the facilities and making the problem tractable.

For our studies, we consider a power generation system consisting of conventional power plants and a wind park which is responsible for supplying a region of certain dimension with electrical energy. With the aim of balancing supply and demand and thus achieving a better capacity utilization of the power plants, also energy storages are integrated into the system. Additionally, we include the possibility of purchasing electrical energy from the spot market. On this basis, the operation of the facilities is optimized with the aim of a cost-efficient generation of energy, see also [4].

More precisely, the optimization criterion of the problem comprises the overall costs incurred by the satisfaction of the consumer demand within the planning horizon. In detail, they consist of the operational costs of the generation units and

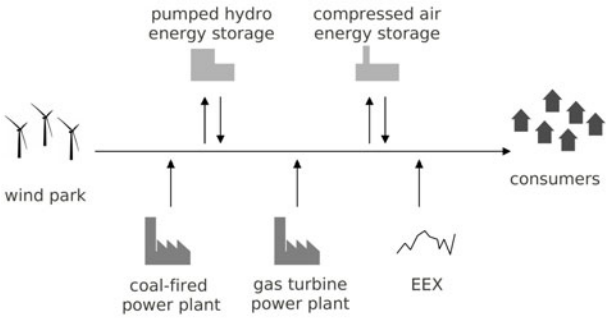


Fig. 1 Schematic power generation system

storages including start up costs of the facilities. Furthermore, the costs of procuring energy from the spot market are taken into account.

One of the major requirements in a generation system is the satisfaction of demand at any time. This means that in each time step of the planning horizon, the demand has to be covered by the total power produced by the power plants, purchased on the spot market, produced from wind power, or obtained from the energy storages. The corresponding constraints occupy a central position in the model, as they couple all facilities of the system.

Concerning the technical description of the facilities, one important aspect regards the modeling of the switching processes which require the integration of binary variables. In particular for hard coal power plants, these variables are also used to pose minimum runtime and downtime restrictions. These constraints are taken into account in order to avoid increased thermal stress of the units.

Furthermore, the consideration of partial load efficiency of the facilities is necessary in order to achieve a realistic description. More precisely, certain characteristic curves are assigned to each facility representing its operational behavior in dependence of the current production level. As the resulting efficiency functions normally behave nonlinearly, we follow the approach of a piecewise linear approximation of these nonlinearities in order to handle the resulting complex problem. Although the approximation leads to an increased number of binary variables and constraints, it enables the solution of problems of larger size, compared to the nonlinear formulation.

A special model requirement is the integration of uncertainty with regard to some unknown parameters. As the wind power production strongly depends on meteorological conditions, the uncertainty concerning the amount of wind energy should be taken into account. Also the market price for electricity is unpredictable and varies over time. In order to generate solutions that hedge against this uncertainty, we formulate a multistage stochastic mixed-integer problem. Here, the uncertainty is described by a multivariate stochastic process which is represented via a scenario tree. Each scenario of the tree corresponds to one realization of the stochastic process to which a certain probability is assigned. By making use of the scenario tree formula-

tion, the problem described above can be formulated as a large scale mixed-integer optimization program. A detailed description of the model is given in [7].

3 A Scenario Tree-Based Decomposition Approach

The solution of multistage stochastic mixed-integer programs still poses a great challenge from the computational point of view, since integer as well as stochastic aspects are comprised in one model. The need for modeling combinatorial decisions combined with uncertainty motivated a number of contributions exemplarily presented in the sequel.

One of the first papers concerning the solution of two-stage stochastic mixed-integer stochastic problems was published by [5] proposing the Integer L-Shaped method. For the solution of two-stage problems with integer variables in both stages, for instance [3] propose a dual decomposition algorithm employing Lagrangian relaxation for the decoupling. Concerning the optimization of multistage stochastic mixed-integer problems, we exemplarily refer to [8, 6, 1]. Most of these contributions show some similarities to our method, as they also use a branch-and-bound approach in combination with a decomposition method. However, in order to solve our application problem, we need to exploit the problem specific characteristics which are specified below. Rather than using a scenario decomposition, our approach is based on the subdivision of the corresponding scenario tree, which will be presented in the following.

By formulating the stochastic problem as described in Section 2, we obtain a large-scale, block-structured mixed-integer problem. Algorithmically, this structure makes the problems attractive for decomposition approaches which exploit this characteristic by splitting the entire problem into manageable subproblems. Especially in the linear case, successful decomposition approaches have been developed, see e.g. [2]. But also for problems including integrality restrictions, decomposition approaches are very promising as indicated above.

Hence, for the solution of the application problem, we have developed a novel decomposition method with the aim of exploiting the problem-specific characteristics. Among others, the following two observations have mainly influenced the development of our decomposition approach. First, the problem shows a loose connectivity with respect to variables associated with different nodes of the scenario tree. In particular, two time steps are only coupled by the storage balance equation, the upper bound on the power gradient and the minimum run time and down time restrictions. The second and more important observation is based on computational investigations regarding the solution of smaller problem instances. Namely, the fixation of variables at a selected node of the scenario tree has only little impact on the optimal solution values of variables associated with nodes which are sufficiently far away. This means that in most cases, the optimal decisions corresponding to a node n are not changed if a variable of a further node m is fixed and the distance of n and m exceeds a certain path length in the scenario tree. Under consideration

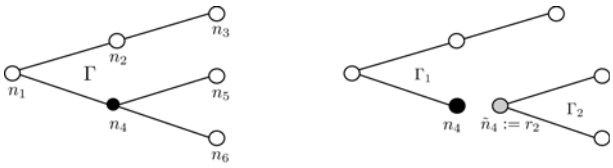


Fig. 2 Exemplary splitting of a scenario tree with 6 nodes

of these facts, our goal is to employ a decomposition which generates subproblems whose coupling between each other is hardly correlated and thus, exploits the lack of sensibility described above.

The basic concept of the developed algorithm takes this decomposition suggestion into account and combines it with a branch-and-bound procedure. The underlying idea is based on the partition of the scenario tree into several smaller subtrees by defining so called *split nodes* where the tree is split up. This procedure is exemplarily illustrated in Figure 2, where the solid node represents the split node. Using this subdivision, the subproblems are formulated independently based on the resulting subtrees. Note that in contrast to the scenario decomposition only variables corresponding to split nodes need to be doubled. The formulations are connected by adding so called *coupling constraints* yielding a reformulation of the original problem. If the coupling constraints are relaxed, the problem decouples into a collection of separate subproblems, which can be solved independently providing a lower bound on the optimal objective function value. In order to ensure feasibility, the decomposition is embedded within a branch-and-bound framework. This means that by branching on pairs of original and duplicated variables, the satisfaction of the coupling constraints is recovered. A detailed description of the scenario tree-based decomposition combined with branch-and-bound, called *SD-BB* algorithm, is given in [7].

The *SD-BB* approach provides the following advantages. Decomposing the problem with respect to predefined split nodes allows us to determine the size of the subproblems depending on the individual problem instance. Here, we remark that the resulting subproblems are still mixed-integer formulations which makes a suitable size desirable in order to achieve a good performance. Furthermore, the algorithm shows the useful characteristic that in each branch-and-bound node at most one subproblem has to be solved in order to obtain a lower bound on the optimal function value. Indeed, subproblems with identical branching bounds may appear various times during the solution process. This fact can be exploited by a suitable caching procedure which stores already solved subproblems in order to avoid redundant solutions of similar problems. Moreover, we can benefit from the flexibility of the branch-and-bound approach concerning the application of further techniques to speed up the algorithm, such as the integration of Lagrangian relaxation, problem specific heuristics, branching strategies, and separation algorithms.

We have evaluated the performance of the *SD-BB* approach based on a series of test runs considering instances of different characteristics, where the number of facilities in the supply system range from two to twenty. Thereupon, we proposed a

general setting of parameters and methods for the application to further instances. Summarizing these results, the *SD-BB* algorithm is able to solve large instances with a time horizon up to four days and a time discretization of 15 minutes to optimality or at least to provide a quality certificate of a relative gap less than 1%. The results obtained by our approach are also compared to the standard commercial solver CPLEX, indicating the suitability of the *SD-BB* for the solution of the application problem.

4 Conclusion and Outlook

Altogether, we have developed a scenario tree-based decomposition approach for the solution of multistage stochastic mixed-integer programs and applied it successfully to the problem described above. Although we conceived the *SD-BB* algorithm for the solution of the energy production problem, its general framework is applicable to a wide range of related problems and lends itself to the solution of problems with similar structural characteristics.

This novel approach offers several aspects for further research and improvement. For instance, a further development of selected routines of the standard branch-and-bound approach such as preprocessing techniques and more elaborate branching rules provide a potential for an improved performance. In particular the handling of continuous splitting variables represents a point for further research, since their number strongly influences the running time.

References

1. A. Alonso-Ayuso, L.F. Escudero, and M.T. Ortuño. BFC, A branch-and-fix coordination algorithmic framework for solving some types of stochastic pure and mixed 0-1 programs. *European Journal of Operational Research*, 151: 503–519, 2003.
2. J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Verlag, 1997.
3. C.C. Carøe and R. Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24: 37–45, 1999.
4. A. Epe, D. Mahlke, A. Martin, H.-J. Wagner, C. Weber, O. Woll, and A. Zelmer. Betriebsoptimierung zur ökonomischen Bewertung von Speichern. In R. Schultz and H.-J. Wagner, editors, *Innovative Modellierung und Optimierung von Energiesystemen*, volume 26 of *Umwelt und Ressourcenökonomik*. LIT Verlag, 2009.
5. G. Laporte and F.V. Louveaux. The integer L-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters*, 13: 133–142, 1993.
6. G. Lulli and S. Sen. A branch-and-price algorithm for multi-stage stochastic integer programming with application to stochastic batch-sizing problems. *Management Science*, 50: 786–796, 2004.
7. D. Mahlke. *A Scenario Tree-Based Decomposition for Solving Multistage Stochastic Programs with Application in Energy Production*. PhD thesis, Technische Universität Darmstadt, 2010.
8. M.P. Nowak and W. Römisch. Stochastic lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty. *Annals of Operations Research*, 100: 251–272, 2000.

Train Scheduling in a Large and Highly Utilised Railway Network

Gabrio Caimi

Abstract This paper proposes a comprehensive approach for the railway scheduling problem, which starts with the commercial description of intended train services and has the goal of generating a conflict-free detailed schedule. The approach consists of three description levels with their corresponding interfaces and enables a hierarchical divide-and-conquer approach. The first step is the description of a formal structure for describing the service intention, including periodicity information. A projection scheme is used to create an augmented periodic problem. This augmented periodic timetabling problem is solved first globally on an aggregated topology and simplified safety model, and subsequently refined locally by considering all details of the infrastructure and train dynamics. Finally, the generated periodic conflict-free schedule is rolled out over the complete day to create a production plan fulfilling all requirements specified in the service intention.

1 Introduction

In this paper we present an integrated system-wide approach for timetable planning, from the description of the commercial offer as a starting point to the conflict-free detailed train schedule (the production plan) as the outcome. For this purpose, a multilevel approach is proposed. It starts with the formalisation of the commercial services that the railway companies would like to offer to the customers over a whole day. Starting with this input, we develop methods and algorithms for reaching the goal of a network-wide detailed conflict-free train schedule for a whole day. We partially use well-known methods from the literature, combined with methods especially developed during this project.

The multilevel planning approach is conceived as a decision support system for the planners. Thus, we strictly separate technical decisions, which ought to be au-

Gabrio Caimi
BLS Netz AG, Genfergasse 11, 3001 Bern, Switzerland e-mail: gabrio.caimi@bls.ch

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings, 659
DOI 10.1007/978-3-642-20009-0_104,

tomatic, from commercially relevant decisions, which should be decided by human planners and given as input. Nevertheless, planners can intervene at each level for manually adjusting the automatically generated schedule if necessary.

The purpose of this paper is to describe the general architecture of the proposed multi-level approach. For more details on the different levels and for further computational results the reader is referred to [1].

2 The Service Intention

A service intention is a precise description of the commercially relevant part of the railway offer, which serves as basis for negotiations between train operating companies and infrastructure managers. These have the task to transform this service intention into a detailed production plan.

Periodic timetables are common in many countries. The demand, however, is not distributed uniformly over the day, and also different days of the week can have different demands. Thus, the resulting timetables are not completely periodic, but typically have a strong periodic structure with the addition of numerous non-periodic exceptions (e.g. additional trains in peak hours). In this paper, the notion of a *partial periodic service intention* is therefore proposed. It consists of sets of the following three elements. First, a *train run* is defined as the run over a line with given *stopping stations*, specifying the minimal and maximal values for the *stopping time* in the stations, and the *trip time* between two consecutive stations, and at least one *time slot* for a departure or arrival event of the first train recurrence. Moreover, it also specifies the *number of repetitions* of the train run and the given *periodicity*. Second, a *connection* is defined as the possibility for the passenger to change from a train to another train in a specified *station*, with given *minimal and maximal time* for the connection to take place. As third element, a *time dependency* is defined as a time constraint between two train runs with *lower and upper bound* for the time difference between the given departure or arrival times.

Given a partial periodic service intention, the task is now to create a production plan that fulfils the specified requirements and is operationally feasible. The approach presented here is based on a reduction of this (partial periodic) input to a particular fully periodic instance, for which algorithms for periodic timetabling can then be applied. The periodic schedule generated in this way will then be rolled out to a schedule for a complete day.

For large railway networks, the very large amount of data makes it impossible to incorporate all details of the railway topology in a single planning step. We therefore follow a two-level approach for the train scheduling problem based on two different description levels of the infrastructure, similar to the DONS project in the Netherlands [3]. The first, *macroscopic level* considers only a simplified version of the topology with the most relevant information, while the second, *microscopic level* considers the detailed topology on a locally bounded region.

3 Macro Scheduling

Train scheduling on the macroscopic level (*macro timetabling*) focuses on global interdependencies over the entire network for generating the most important properties of the timetable and thus has to avoid dealing with large amount of detailed information that is only locally relevant. According to the simplified topology model, also the train movements and train dynamics are simplified, basically by taking only the travel time between two consecutive stations into consideration.

A macro timetable consists of a set of event times for departure and arrival of trains at the stations. The event time is here denoted by π_i for an event number i, with periodicity T. The choices of the event times are dependent through requirements of trip times, dwell times and connections. These dependencies can be expressed via inequalities representing the periodic event time difference between events i and j. This problem formulation is well known as the Periodic Event Scheduling Problem (PESP) [4].

The *Flexible Periodic Event Scheduling Problem* (FPESP) is a generalised version of the PESP that is well suited for the macro scheduling within the presented multilevel scheme. It takes into account that the solution of the macro level still is to be refined and checked in the micro level. The goal is to produce a flexible output from the macro level to have a larger solution space in the micro level. This way, the chance to get a feasible solution on the micro level are substantially improved, still remaining not guaranteed.

Instead of determining event time points, the idea is to generate time slots $(\pi_i, \pi_i + \delta_i)$ for the events, where δ_i is the size of the flexibility range. Such flexible time slots are shown in Figure 1.

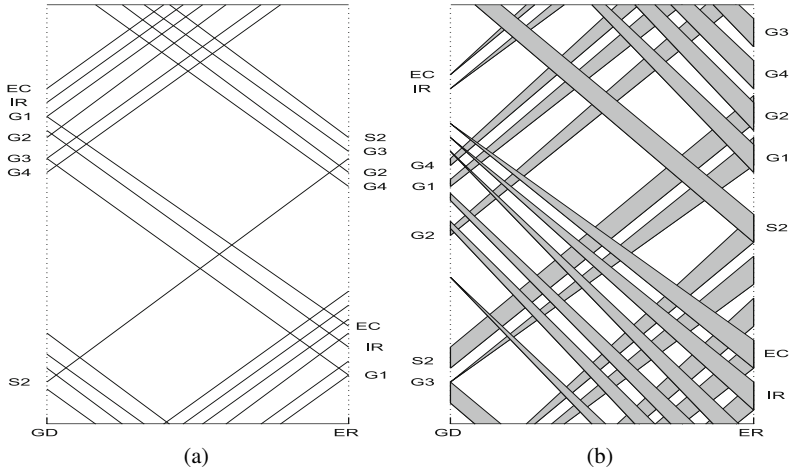


Fig. 1 (a) The generated timetable without event slots visualised in a time-space diagram. (b) When using event slots, each event gets an event time π_i and a flexibility δ_i.

Fulfilling the PESP-constraints for every combination of event times in the intervals is equivalent to requiring

$$l_{ij} + \delta_i \leq \pi_j - \pi_i + T p_{ij} \leq u_{ij} - \delta_j \quad \forall\, (i,j), \tag{1}$$

where l_{ij} , u_{ij} are the lower and upper bound of the corresponding PESP-constraint.

These FPESP constraints (1) are very similar to the PESP constraints, with the additional continuous decision variables δ . The well known MIP formulations of the PESP model can be easily adapted. The presented FPESP extension makes sure that the data transmitted between macro and micro level is not more restrictive than necessary, leading to a larger solution space for the latter.

The amount of generated flexibility as well as the CPU time can vary considerably. With an appropriate choice of the objective function, for our test instance in central Switzerland (1083 variables, of which 436 integer, and 1730 constraints) it is possible to generate in reasonable time (57 seconds instead of 14 when computed without flexibility) a timetable with a total flexibility of over 180 minutes and a comparable timetable quality from the commercial point of view.

4 Micro Scheduling

The existence of an operable production plan for a given draft timetable has to be checked on the micro level (*micro scheduling*) by taking into account detailed information, in particular the exact track paths and their blocking times for ensuring the schedule to be conflict-free.

Micro scheduling cannot be done on the whole network simultaneously because the huge amount of data would lead to both memory problems and prohibitively long computation times. To overcome this problem the network is divided into different zones. For a suitable subdivision, we exploit the networks' typical, very heterogeneous structure and distinguish two types of zones:

- In *condensation zones*, which are usually situated around major stations, the available capacity is scarce. There are many switches and thus overtakings and train crossings are possible and quite common. Because of this structural complexity and high traffic density, condensation zones are usually the bottlenecks of the network.
- *Compensation zones* connect the condensation zones and have typically more capacity available as well as a much simpler topology (some parallel tracks with few crossings and switches).

The most critical zones to schedule are the condensation zones. There, the choice of an appropriate routing is the crucial degree of freedom to exploit for finding a conflict-free train schedule. For this task the *conflict graph model* [5] has become one standard approach in the literature. Unfortunately, the LP-relaxation of this problem is very weak, such that large instances in this model can only be solved heuristically or with time-consuming pre-processing.

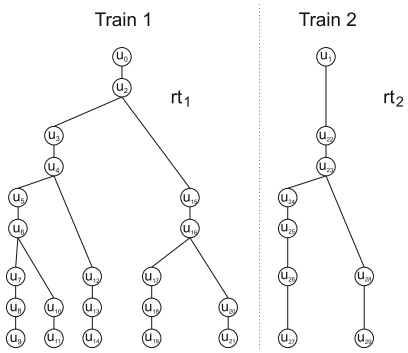


Fig. 2 Example of Resource Trees for two trains.

It is possible to characterise the set of conflict-free routes by a different model in a much more concise way [2]. Instead of a simple binary relation between the nodes representing complete train routes, we consider the possible routes of a train as a flow through the graph of the single block sections (the resources). This can be represented as a tree structure (the *Resource Tree*), where each branch in the Resource Tree corresponds to a switch where both diverging directions lead to the designated target for the train (the platform or the exit portal), as shown in Figure 2.

The time slots that were generated on the macro level can now be used on the micro level to generate a (discrete) set of alternative arrival and departure times inside the given time interval. It is therefore necessary not only to assign a route to each train, but also to choose one of these alternative times for each arriving and departing train.

Afterwards, each edge in the Resource Tree (i.e., each used topology element) is directly associated with the corresponding resource in the railway topology indicating its blocking time interval. These blocking times are computed precisely by taking into account the detailed topology and signal information. Given all these blocking times, it is possible to compute very efficiently the so-called *conflict cliques*, which limit the concurrent allocations of a resource to at most one train, separately for each resource. Finally, the set of alternative scheduling possibilities for each train and the detected conflict groups can be merged to a multi-commodity flow formulation, where for each train a path through its Resource Tree (one integer flow unit from the root to one of the leafs) is sought that meets the conflict constraints. We call this the *Resource Tree Conflict Graph* (RTCG) model.

The RTCG can be formulated as an integer linear program (ILP) that can then be solved with a commercial solver. The improvement in structure achieved with the presented RTCG model is so strong that in many cases the LP-relaxation was already integer, and realistic, large-scale instances could be solved in a few seconds. In our test case, the timetable obtained from the macro level gives rise to 1217 different routes in the condensation zone Lucerne, which results in an RTCG with 44'327 nodes, 43'487 blocking edges, and 43'517 flow edges. The resulting MIP has 43'517 binary variables, but only 1'999 conflict constraints. After 10 seconds

of pre-processing, the MIP solver was able to deliver a relaxed solution that was already integer.

5 Conclusion and Outlook

In this paper, a divide-and-conquer approach has been presented for conflict-free timetabling in railway network, in particular for dense services with heterogeneous periodicity. The presented multilevel strategy and the developed models for macro and micro scheduling allow planning a system-wide train schedule, from the description of the commercial offer to the detailed production plan. It can be of great help to the planners for strategic, tactical, and short-term planning. For this goal, we have adapted and applied known models and algorithms for each level, based on the state-of-the-art from the literature and our own previous and current work.

However, with this milestone we are still only at the beginning for a model-based co-production of planning and operation. The real challenge comes with the introduction of a decision support system for dispatching. With the topic of rescheduling, i.e., the continuous adaptation of the current production plan to the real-time situation, a very interesting research area will be opened, which will set new limits for the current Operations Research models while requiring an even stronger collaboration between the involved actors in planning, operation and process design.

Acknowledgements The author thanks the Swiss Federal Railways, Infrastructure Division for funding and providing data as well as his former colleagues at ETH Zurich for their fundamental support.

References

1. G. Caimi. *Algorithmic decision support for train scheduling in a large and highly utilised railway network*. PhD thesis, ETH Zurich, 2009.
2. G. Caimi, F. Chudak, M. Fuchsberger, M. Laumanns, and R. Zenklusen. A new resource-constrained multicommodity flow model for conflict-free train routing and scheduling. *Transportation Science*, 2010. Article in Advance.
3. L. G. Kroon, D. Huisman, E. Abbink, P.-J. Fioole, M. Fischetti, G. Maroti, A. Schrijver, A. Steenbeek, and R. Ybema. The New Dutch Timetable: The OR Revolution. *INTERFACES*, 39(1): 6–17, 2009.
4. P. Serafini and W. Ukovich. A mathematical model for periodic scheduling problems. *SIAM J. Disc. Math.*, 2(4): 550–581, 1989.
5. P. J. Zwaneveld, L. G. Kroon, H. E. Romeijn, M. Salomon, S. Dauzère-Pérès, S. P. M. van Hoesel, and H. W. Ambergen. Routing Trains through Railway Stations: Model Formulation and Algorithms. *Transportation Science*, 30(3): 181–194, August 1996.

Author Index

B. Hu et al. (eds.), *Operations Research Proceedings 2010*, Operations Researrch Proceedings,
DOI 10.1007/978-3-642-20009-0_104, © Springer-Verlag Berlin Heidelberg, 2011